Computational Analysis of Transporter Repertoires as Determinants of Cellular Cancer Drug Response

Pisanu Buphamalai

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Aachen 29.07.2016

Thesis supervisors:

Prof. Juho Rousu

Thesis advisors:

Prof. Julio Saez-Rodriguez

Dr. Marc Brehme

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Pisanu Buphamalai

Title: Computational Analysis of Transporter Repertoires as Determinants of Cellular Cancer Drug Response

Date: 29.07.2016 Language: English Number of pages: 6+44

Department of Computer Science

Professorship:

Supervisor: Prof. Juho Rousu

Advisors: Prof. Julio Saez-Rodriguez, Dr. Marc Brehme

Cancer patient heterogeneities challenge disease management despite advances in targeted therapies. Patient sub-populations are irresponsive to certain treatments with unknown reason or differ in sensitivity, while mutations can cause resistance and patient relapse. Several consortia of large-scale pharmacogenomic screens against comprehensive panels of human cancer cell lines have therefore been established. Although these resources provide novel genetic biomarkers of cancer drug response, the predictive accuracy has still been lower than desired. It is hypothesised that many drugs act upon their endogenous targets by hitch-hiking on membrane channels. About 10% of the human genome encodes for transport-related functions, a functional link between transporters and disease relevance are yet largely unknown. This thesis therefore focuses on the roles of membrane transporters in cellular drug response. Solute Carriers (SLCs), which represent the second-largest family of membrane proteins in the human genome and the largest class of transporters, are the central focus of this study. For systematic identification, the analyses also include more well-known transporter family of ATP-binding cassettes (ABC), which have a widely accepted role in mediating drug resistance. The landscape of expression for both transporter families in cancer tissues were analysed. Matrix factorisation based methods were also employed in integrating multiple genetic data to observe tissue specificity patterns. It was found that differential expression of several transporters are likely to link with tumorigenesis. Moreover, functions of transporters in drug influx and efflux were also computationally hypothesised. Several statistical methods were used and compared, with a list of most potential candidates suggested for experimental validation. The interaction between two transporters were also identified using linear model with interaction terms. Both computational and biological challenges and limitations of the project are discussed.

Keywords: Solute carriers, transporter, drug sensitivity, interaction learning

Preface

This report presents the work that I have done during my master's thesis, which is a part of the Erasmus Mundus Master's Programme in Systems Biology (euSYSBIO), jointly supervised by KTH Royal Institute of Technology in Sweden and Aalto University in Finland. The project is conducted in the Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University. I would like to take this opportunity to express my gratitude towards several people that have helped me throughout my master's studies.

I would firstly like to thank Prof.Julio Saez-Rodriguez and Dr. Marc Brehme, who accepted me to join the group and this had opened up my knowledge in the field of systems pharmacogenomics, and gave me flexibility of exploring relevant methods with my full capability to exploit in the project. I would also like to thank all the members of JRC-COMBINE, who have been very helpful in giving new ideas and feedback, especially Mi Yang who works alongside in the same project, and Jorge Ferreira for fruitful discussion regarding the biology of transporters in metabolic network.

Moreover, I would like to thank Prof. Samuel Kaski and Dr. Pekka Marttinen of Aalto University, who supervised me during my summer internship. Machine learning has been a valuable tool in computational biology and has been exploited in this project. I would also like to thank Dr. Nathan Harmston, my first research mentor, who taught me several concepts in computational biology during my undergraduate studies.

I would like to express my gratitude towards Prof. Erik Aurell and Prof. Juho Rousu, for their help of supervising and coordinating this master's thesis.

And lastly, I would like to thank my family and friends who always have been of a great support to me.

Aachen, July 2016

Pisanu Buphamalai

Contents

Operators and abbreviations

Operators

$\text{tr}(\mathbf{X})$	trace of a matrix $\mathbf{X}$, i.e. sum of its diagonal elements
$\|\|Z\|\|_p$	ℓ^p norm of a vector Z
$\mathbf{X}'$	Transpose of a matrix $\mathbf{X}$
$\sum_i$	sum over index i
$\mathbf{A} \cdot \mathbf{B}$	dot product of vectors $\mathbf{A}$ and $\mathbf{B}$

Abbreviations

ABC	ATP-binding cassette membrane transporter superfamily
ANOVA	Analysis of variance
AOC	The scaled area over drug response curve (AOC $\in [0, 1]$)
CCLE	The Cancer Cell Line Encyclopedia
CFE	Cancer functional events
EN	Elastic Net regularisation
IC_{50}	The concentration of an inhibitor where the response (or binding) is reduced by half
GDSC	The Genomics of Drug Sensitivity in Cancer project
SLC	Solute carrier membrane transporter superfamily
TCGA	The Cancer Genome Atlas

1 Introduction

Drug discovery is a costly process. The advancement of genome sequencing technology has allowed scientists to detect the genetic causality of several diseases and this has been an important process for identifying putative drug targets and discovery. However, the average number of FDA approved drugs per year has fallen since 1990s [1] despite human genetic information was more well understood. It has also been estimated that the cost for pharmaceutical R&Ds will be doubled every 9 years [2]. One of the reasons that the vast majority of investigational drugs failed during clinical trial is that they demonstrated lower efficacy than expected while another reason is insufficient safety as drugs may develop toxic side effects during the trials [3]. Both of these issues can be hypothesised to be associated with functional roles of membrane transporters in cellular drug disposition.

The success of drug disposition can generally be determined by four criteria: absorption, distribution, metabolism, and excretion (ADME). Absorption refers to the uptake of the drug in the body, normally through oral route. Distribution describes specific tissues or organ where the drug appears to accumulate after absorption. Metabolism refers to the effect of the drug that induces chemical changes via enzyme catalysis, and the drug finally will be eliminated usually by the kidney or liver as urine or faeces. Several steps in ADME criteria are associated with cellular influx and efflux of drugs via membrane transporters. It is therefore important to understand mechanisms of drugs entering a cell through membrane transporters, yet this have mostly been understudied.

In this thesis, the main focus is therefore on the systematic elucidation of the role of the two largest superfamilies of transporters, solute carriers (SLC) and ATP-binding cassettes (ABC) transporters as determinants of cellular response to cancer drugs. The two superfamilies have been evidenced to play more significant roles in cellular drug uptake than other transporters or membrane channels and hence are the central focus to this study. Overview of other transporters are however briefly discussed in Chapter 2.

In Chapter 2, previous studies related to transporters are discussed. The endogenous functions of SLC and ABC transporters have been linked to tumorigenesis, which can be a result of insufficient or excessive intake of metabolites that facilitate cell growth [4]. The discovery of high association of particular transporters in drug sensitivity are also discussed here. Transporters such as ABCB1 is known as multidrug resistance transporter (MDR) due to its ability to utilise ATP hydrolysis to actively pump several small molecules outside the cell [5]. On the contrary, SLC transporters are more regarded as influx transporters [6]. They transport small molecules including drugs that structurally resemble the natural metabolites of the transporter ('metabolite-likeness' [7]). However, SLCs are currently regarded as the most neglected transporter family in human [8].

Large-scale pharmacogenomic screening in cell lines has been a valuable database for discovering novel drug targets. Drug screening consortia such as the Genomics of Drug Sensitivity in Cancer (GDSC) Project [9] provide valuable data in systematic identification of genetic effects in drug sensitivity. The main pipelines from GDSC

consortium and the consistency of the data across multiple consortia [10] were also discussed in Chapter 2.

Cellular transporters and their association with cancer drug mode of action have not been characterised systematically [8], and this study therefore aims to address this challenge. Genomic information across 1001 cell lines with drug sensitivity data across 265 drugs were obtained from GDSC consortium, and several methods were used to (1) identify the tissue characteristic expression specificity of all SLC and ABC transporters using the moderated (Bayesian) *t*-test [11], (2) investigate the transporter similarity of different cancer tissues, using matrix factorisation method `iCluster` [12, 13] to integrate multiple genomic profiles, (3) detect the association of transport gene expression level to drug sensitivity using multiple methods i.e. correlation test, ANOVA, and linear model with regularisation, and (4) identify the dependency for drug intake between two transporters using group-lasso based linear model [14]. The methods and research materials are provided in Chapter 3.

Chapter 4 provides results from the analyses mentioned above. In the landscape analysis, transporters that are over-expressed or under-expressed in particular tissues were identified, and the strongest associations are linked to literature evidence for cancer progression. In drug association analysis, several known and novel associations were captured by different methods tested. The computational and biological limitations and challenges of these results are discussed. Furthermore, it is promising to investigate whether there are particular pairs of transporters that function together in drug intake and uptake. Linear model with regularisation with interaction terms of the features were included and a number of interactions were captured.

The combination of association results were discussed in Chapter 5 in order to serve as rationale for candidate selection and prioritisation for experimental validation of the top hits. A number of suggestions for improvements and the possible future directions of the project were also included in this chapter, with limitations and drawbacks of the methods used in the analyses also being discussed.

2 Background

2.1 Membrane Transporters

Membrane transporters serve as gateways for the exchange of endogenous and exogenous substances between different transmembrane compartments to control the cellular homoeostasis. While some molecules can diffuse through cell membranes, many of them are lipid insoluble and require a transporter to enter the cell [7]. The expression patterns of transporters are greatly varied, many of them are expressed throughout multiple tissues, especially in epithelial tissues such as liver, kidney, and intestine, and those with barrier functions [15]. Transporters can be found in plasma membrane, and for eukaryotic cells, other membrane-bound organelles such as mitochondria and vesicles. Transporters can be classified into solute carriers (SLC), ion and water channels, and ATP-driven transporters or pumps. Two of the largest classes of transporter families are solute carriers (SLC) and ATP-binding cassette containing transporters (ABC) have been found to play critical roles in transporting drugs inside the cells besides their major roles in transporting small molecules such as nutrients and metabolites [5] [15]. Strong association between drug efficacies and the abundance of particular transporters have been reported, leading to more attention towards these two superfamilies. However, most studies of transporters were performed in a small-scale experiment, particularly focusing on a number of transporters to some drugs. Little evidence of studies focusing on those transporters in genome-wide level, i.e. how the transporter of interest perform compared to other transporters, and whether there are more complex factors associated drug uptake. This holistic view of analysis will be referred to as systematic studies.

Other membrane transporters and channels that mediate specific substrates include water channels (aquaporin) and ion channels. There are fewer evidences about these channels in mediating drug uptake, which may be explained by the structural dissimilarity of drugs and the substrates of these channels (water and ions). The overview of membrane transporters and their localisations are displayed in Figure 1.

2.1.1 Solute Carriers (SLC)

Health implications

Solute carriers (SLC) is the largest membrane transporter superfamily, consisting of over 400 genes in the human genome [8]. It is the second largest family of membrane-coded genes in the human genome. SLC superfamily transporters have major roles in uptake of small molecules inside the cell. Different transporters are specific to certain molecules (substrate), resulting in various specificity range to drugs. For instance, transporters with broad range of substrate specificity such as organic anion transporter 1 (OAT1 or SLC22A6), which naturally associate with the transport of various substrates, have consequently been associated to several drugs. Unfortunately, SLC transporters with narrower range of substrate specificities (those that naturally carry only fewer substrates) have received very little attention

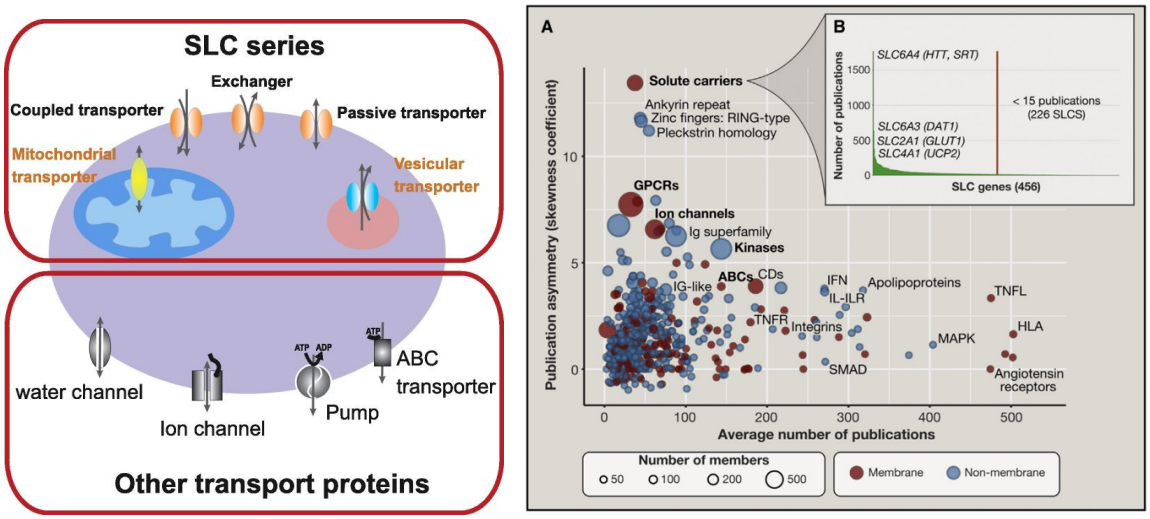


Figure 1: *Left*: Figure from Hediger *et al.* [16] showing types of membrane transporters in human. The upper rectangle shows the superfamily of solute carrier (SLC series) which can be localised in the plasma membrane or other organelles. The lower rectangle shows other types of transporters including water channel, ion channels and pumps. Pumps require energy from ATP hydrolysis in efflux of molecules from intracellular space (lower concentration) to the extracellular environment (higher concentration of substances).
Right: the publication asymmetry from [8]. SLC is a relatively large family compared to other protein families, yet received little attention compared to other protein families, according to the number of publications.

from the community [17]. However, these transporters can be highly associated to particular drugs. The impact of those narrow substrate SLC transporters to health has been indicated through knockout studies of particular transporters as well as the phenotypes observed for patients with Mendelian diseases. Lin *et al.* [17] have recently reviewed the impact of SLC transporters in monogenic disorders and also more common disorders through genome-wide association studies (GWASs). Information regarding SLC transporters substrate specificity have been collected in www.bioparadigms.org. The database contains brief information about their natural substrates as well as associated diseases. The information regarding drug specificity, as previously mentioned, is barely known for the majority of transporters.

The endogenous functions of SLC transporters at a physiological level have been highlighted [15], suggesting that the SLC and ABC superfamilies are likely to be a part of the complex signalling processes known as remote communication system in different scales. Nigam [15] explained it as the effect of transporters, especially those in tissues that are interfacing body fluids, in controlling the exchanges of small molecules involved in signalling pathways or in key steps of metabolism. This sensing mechanism could be found at intra- and intercellular level, and are believed to be present at tissue and organ level. This findings has led to higher complexity level of indirect effects of transporter functions in diseases.

Despite the importance of SLC transporters, they are the most neglected group of genes in the human genome. César-Razquin *et al.* [8] calculated the 'publication asymmetry' (Figure 1), which is measured by the skewness of the distribution of the number of publications for all genes in a group. There is only a small proportion of genes in the SLC superfamily that have been well studied with significantly high number of publications (SLC6A4, serotonin transporter gene has been mentioned in over 1,500 publications) while the majority of genes in the superfamily have not yet been caught sufficient attention in the scientific community (226 SLC genes have been mentioned in less than 15 publications). This is therefore a challenge for systematic characterisation of SLC families, especially those transporter with lesser known functions.

SLCs as drug transporters

As briefly discussed in Chapter 1 that a drug can be effective when it successfully enters the cell and reaches its target while being disposed in the right amount. It is therefore important to understand mechanisms of drugs entering a cell through membrane transporters, yet this have mostly been understudied.

Besides the endogenous roles for the influx and efflux of metabolites and nutrients, many SLC transporters appear to be 'multi-specific drug transporters'. Kell [7] and Dobson & Kell [6] argued that drugs might be able to pass through particular transporters because of their structural similarity to the natural endogenous substrates of the transporters ('hitchhiking'). For example, the knockout of SLC35F2 transporter significantly reduced the efficacy of anticancer drug YM155 [18] [19]. The term 'metabolite-likeness' [20] has been introduced to indicate the importance of the similarity of drug molecules to endogenous substrates.

Many drugs are highly associated with a particular transporter. Additional to transporter SLC35F2 and drug YM155 mentioned above, several other SLC transporters, especially the genes in SLC6 and SLC22 families, have already been targeted by FDA-approved drugs [8]. One major clinical concern regarding drug specificity in this case are potential drug-drug interactions. At transporter level, drugs might compete against each other to transit through the transporter in order to enter the cell. This may lead to drug accumulation in serum or tissues and might induce toxicity. On the contrary, a number of SLC transporters have overlapping substrate specificity which leads to common associations to the same drug. However, due to differential expression of these transporters in different tissues, the specificities in tissue level are varied. Being able to characterise the specificity of theses drugs can therefore be valuable in further identification of drug dispositions (how drug successfully reach the endogenous targets) and estimating drug efficacy and optimal dosage.

2.1.2 ATP-binding cassettes (ABC) transporters

The ATP-binding cassette (ABC) superfamily of transporters comprises about 50 transporter coding genes, which makes it a much smaller superfamily compared to SLCs. They however have important functions related to drug transport [5] [21].

The structure of ABC contains subunits that allow them to utilise energy from ATP hydrolysis to facilitate the transport of various substrates. The roles of ABC transporters have been identified to be primarily associated with the uptake or export of endogenous molecules including several drugs, resulting in drug resistance. A number of ABC transporters have been well characterised, such as ABCB1. It is an important transporter capable of pumping broad range of foreign substrates out of cell and have been found to mediate drug resistance. ABCB1 is also known as multi-drug resistance 1 (MDR1) protein. The other well known ABC transporters that found to be associated with drug resistance are the ABCC subfamily (multidrug resistance associated proteins; MRPs) and ABCG2 [5].

Furthermore, similar to the hypothesis Nigam [15] proposed regarding the endogenous roles of SLC transporters, Fletcher *et al.* [21] also pointed attention to more fundamental roles of ABC transporters in tumour biology, especially their functions in coordinating signalling pathways during tumorigenesis. Targeted functional analyses such as gene knock-out or case-control studies are required to unveil whether ABC transporter expression is associated with cancer progression. Further concerns for future therapies towards ABC transporters, however, are the fact that many of them are functionally redundant. Therefore, combinations of different inhibitors (drugs) will be required to block the transport function more effectively compared to single inhibitor.

2.2 High-throughput small molecule *in-vitro* screening

Understanding molecular mechanisms of cancer and discovery of effective therapeutics remains a challenging task and it has been progressing relatively slowly[3]. As a result of recent biotechnological advancement and availability of high quality genomic data, exploiting the approaches of large-scale data analysis has become possible to overcome the hurdle. Performing robust analyses on pharmacogenomic data at preclinical level could potentially shed lights on the discovery of novel drug targets and consequently reduce the cost and time spent on clinical trials. A number of consortia have been established to specifically study the association of genetic and pharmaceutical information, namely the Cancer Cell Line Encyclopedia (CCLE) [22] and The Genomics of Drug Sensitivity in Cancer (GDSC) Project [23]. Both consortia have collected the genetic information (gene expression, mutation and copy number) of 639 cancer cell lines along with their response in 130 anticancer drugs to investigate their pharmacogenomic interactions. Both consortia used different sources of samples with a proportion of overlapping cell lines. The results of both consortia were compared in Stransky *et al.* [10] leading to the conclusion that their pharmacogenetic data are fairly consistent despite biological and methodological factors that may have caused inconsistency. However, data from both resources are deemed reliable and have been widely used.

Recently, Iorio *et al.* [9] have investigated gene-drug interactions in a larger scale as an improvement of the study in [23], with 1,001 cancer cell lines derived from 29 cancer tissue types and 265 drugs. Different clinically relevant genetic characteristics of all cell lines, referred to as cancer functional events (CFEs), were

extracted from gene expression, copy number, and DNA methylation data. With this information, the authors compared cancer cell lines with primary (patient-derived) tumours, finding that a large proportion of cancer cell lines recapitulate the genomic alterations in patient tumours. Furthermore, the authors also suggest best predictors that associate with drug sensitivity using various methods, and built a logic model to observe the possible combinations of genetic alterations to drug sensitivity.

The 265 anticancer compounds are largely targeting particular pathways, while a small proportion (19 compounds, 7.2%) exhibit cytotoxic effects (inhibiting DNA replication or cytoskeleton) while the rest are targeted agents. These 265 compounds can be categorised in three stages of clinical development, i.e. clinical drugs (n=48), in clinical trials (n=76), and experimental drugs (n=141). Each compound was screened by recording the number of cell population survived at different drug concentrations. The sensitivity was measured by either (1) IC_{50} (the drug concentration of an inhibitor where the response is reduced by half) or (2) the area under the sensitivity curve (AUC). If a drug is sensitive, the cell population would descend in a faster rate and results in lower AUC. The data and results were collected to construct a pharmacogenomics resource, which is now available as the COSMIC and Genomics of Drug Sensitivity in Cancer Web Portal (www.cancerrxgene.org) and is known to be the largest portal for cancer cell lines characterisation to date. The summarising figure of the screened compounds are shown below.

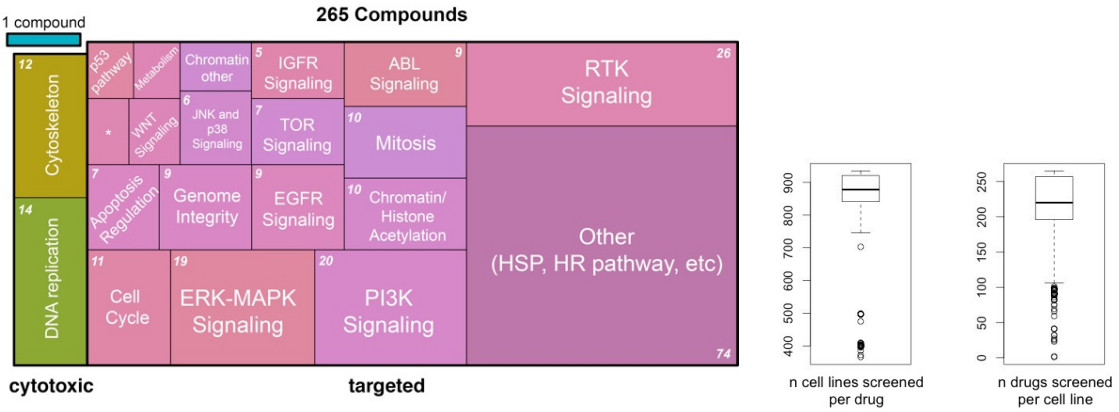


Figure 2: *Left*: The diagram summarising the targets of screened compounds from Iorio *et al.* [9] showing The majority of drugs are targeted in different pathways. *Right*: The boxplot shows the number of cell lines screened per compound and vice versa.

Most anticancer drugs in the GDSC project are majorly targeted, *i.e.* the drugs perturb a specific signalling pathway. However, a small proportion of GDSC drugs exhibit cytotoxic properties (either disrupting DNA replication or cytoskeleton formation). Figure 2 summarises GDSC drug targets. Furthermore, the diagram summarising the analysis procedures from Iorio *et al.* [9] is shown in Figure 3.

With this information, the hypothesis of this thesis is to identify the association between functional roles of transporters and the drug response. The hypothesis is

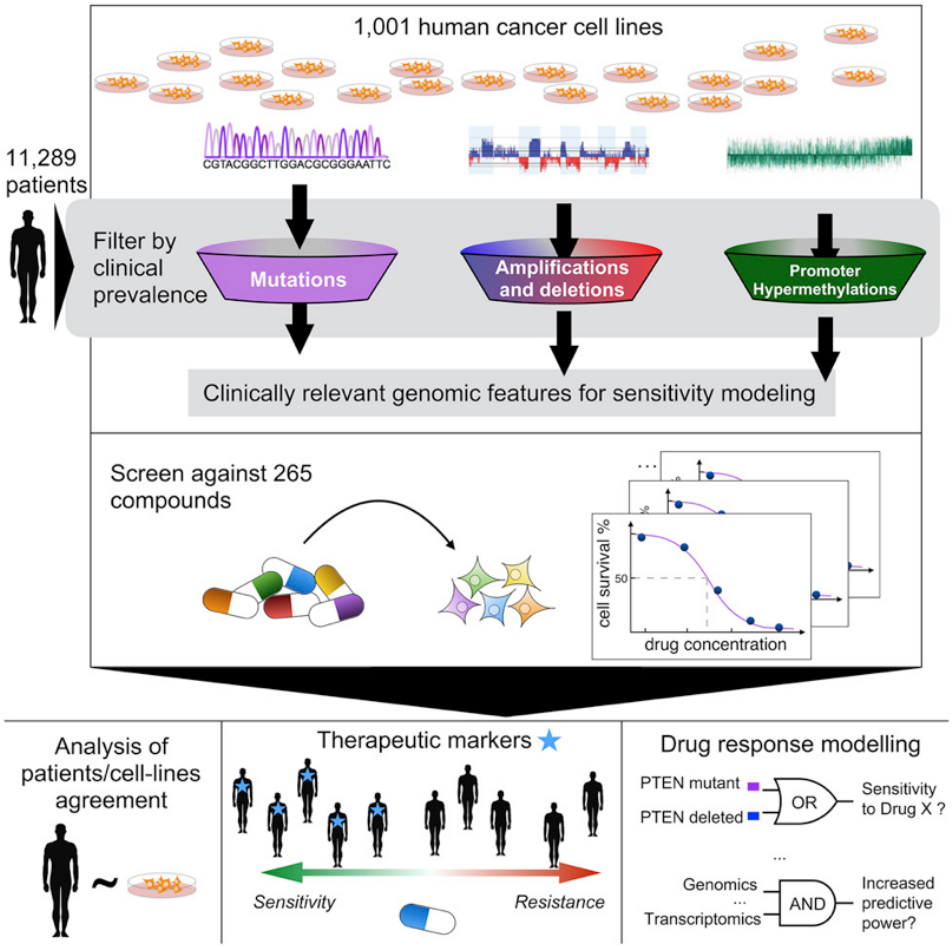


Figure 3: The overview of the analysis pipelines in Iorio *et al.* [9].
Top panel: Three sources of genetic alterations, i.e. mutations, amplifications and deletions, and promoter hypermethylation data from 1,001 cancer cell lines were filtered to capture clinically relevant features. Those filtered data were used as inputs for discovering the features that associate with drug sensitivity.
Middle panel: The 1,001 cell lines were screened against 265 compounds. The sensitivity for each gene-cell line pair was measured using sensitivity curves, which can be represented through IC_{50} or AUC values.
Bottom panel: Associations between genetic profiles and drug sensitivity were identified using different models.

illustrated in the Figure 4:

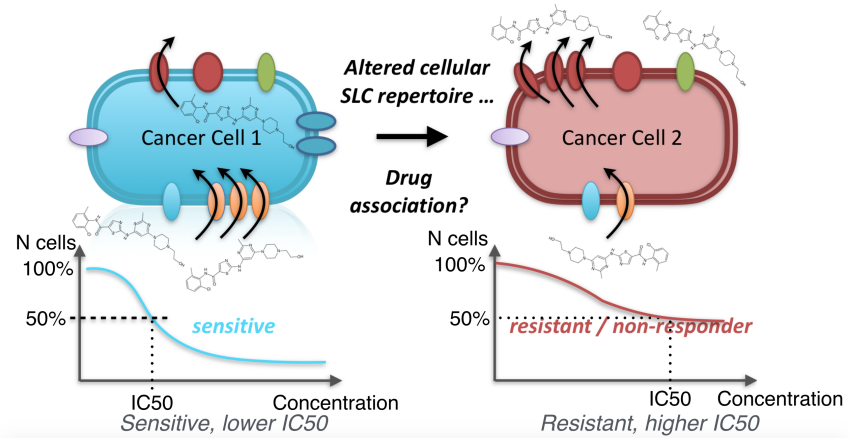


Figure 4: The schematic diagram shows how the effects of influx and efflux transporters can be observed from drug sensitivity curve. With the up-regulation of influx transporters, the drug is more likely to enter the cell and reach its target. This results in high sensitivity of the drug (low IC_{50}, low AUC, high AOC). In the opposite scenario, if the efflux transporters are up-regulated, the drug is more likely to be pumped out of the cell, resulting in drug resistance (high IC_{50}, high AUC, low AOC). However, the up- or down-regulation of a transporter can be directly associated with the metabolic function (as metabolite transporters), and this can also result in similar effect. [Figure courtesy of Marc Brehme]

3 Research material and methods

3.1 GDSC genomic profiles and drug sensitivity data

Gene expression data

The gene expression data is retrieved from RMA-normalised high quality microarray experiments. The expression is stored in a matrix of 978 cell lines by 17,419 genes.

To be able to compare the analyses results in tissue level with primary tumour data from The Cancer Genome Atlas (TCGA)[1], cell lines that tissue labels were unable to identify had been removed from the analysis. The details are shown in the table below:

Table 1: GDSC Cell lines availability based on TCGA tissue labels. Freq columns refer to the number of cell lines belonging to each tissue. Note that there are 812 cell lines (out of 1001, 81.2%) being labelled.

Label	Description	Freq	Label	Description	Freq
ACC	Adrenocortical carcinoma	1	LIHC	Liver hepatocellular carcinoma	17
ALL	Acute lymphoblastic leukemia	26	LUAD	Lung adenocarcinoma	64
BLCA	Bladder Urothelial Carcinoma	19	LUSC	Lung squamous cell carcinoma	15
BRCA	Breast invasive carcinoma	51	MB	Melanoblastoma	4
CESC	Cervical squamous cell carcinoma	14	MESO	Mesothelioma	21
CLL	Chronic lymphocytic leukemia	3	MM	Multiple myeloma	18
COREAD	Colorectal adenocarcinoma	51	NB	Neuroblastoma	32
DLBC	Diffuse Large B-cell Lymphoma	35	OV	Ovarian serous cystadenocarcinoma	34
ESCA	Esophageal carcinoma	35	PAAD	Pancreatic adenocarcinoma	30
GBM	Glioblastoma multiforme	36	PRAD	Prostate adenocarcinoma	6
HNSC	Head and Neck squamous cell carcinoma	42	SCLC	Small cell lung cancer	66
KIRC	Kidney renal clear cell carcinoma	32	SKCM	Skin Cutaneous Melanoma	55
LAML	Acute Myeloid Leukemia	28	STAD	Stomach adenocarcinoma	25
LCML	Chronic Myelogenous Leukemia	10	THCA	Thyroid carcinoma	16
LGG	Brain Lower Grade Glioma	17	UCEC	Uterine Corpus Endometrial Carcinoma	9

Mutation data

The mutation data were taken from the GDSC unfiltered variants catalogue. The data were collected from the same cell lines as gene expression data. The mutation data for 416 SLC and ABC gene were extracted. As a result, 17,207 mutation loci were found.

The mutation is likely random if it was found in only one or a few cell lines. A threshold of minimum 3 cell lines were set for the loci to be used as an input for the analyses. This reduces the number of loci to 140.

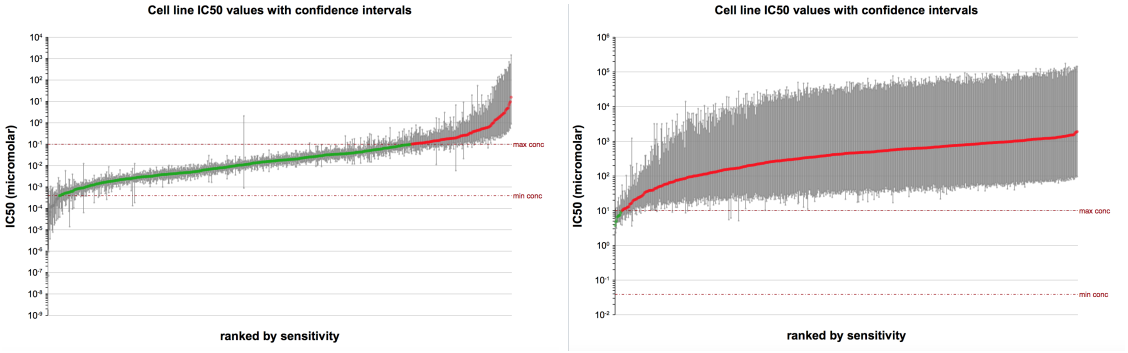
Drug sensitivity data

265 compounds were screened. The number of cell lines screened for each drug are various and was represented in Figure 11 (min = 264, median = 692, and max = 746 cell lines). Each drug has different sensitive concentration ranges, resulting in

[1]Now a part of the Genomic Data Commons (GDC). https://gdc.nci.nih.gov

diverse IC_{50} values, which can be difficult to compare. The area under the sensitivity curve (AUC) was therefore preferentially used. AUC values were normalised to the range between 0 and 1. However, AUC values are not intuitive to use (low AUC refers to high sensitivity), the term 'area over the curve' (AOC) which is defined as 1-AUC will therefore be used from this point, i.e. low AOC value indicates that cell lines are less sensitive to the drug.

Some drugs exhibit low response in the majority of cell lines. An example of the comparison between two drugs are shown below. For quality control in tissue-specific analysis (described in Section 3.4), these drugs were filtered out from the analysis.



(a) Camptothecin (b) GDC-0449

Figure 5: IC_{50} comparison of Camptothecin (left) and GDC-0449 (right). These drug sensitivity plots can be retrieved from the GDSC web portal. The plots show that the majority of the cell lines are mostly insensitive to GSC-0449 and the drug is removed from the tissue-specific analysis.

3.2 Detection of tissue specificity enrichment of SLC and ABC genes

Expression level of transporter genes across multiple cancer tissues can provide insights on the roles of the transporter to the disease progression. This section briefly describes the method used for differential expression enrichment analysis for transporter gene.

Tissue specificity enrichment analysis of a gene is calculated through the *moderated t-test* in `Limma` package in *R* [11]. The test was performed iteratively for each gene and each tissue. For each iteration loop, the expression of selected genes were labelled as 1 if they were from the tissue of interest, and 0 otherwise. `Limma` performs hypothesis testing and reports log fold changes, standard errors, *t*-statistics and *p*-values for significance analysis. The reason of using the moderated *t*-statistics instead of ordinary *t*-statistics have been profoundly discussed in Phipson *et al.* [24] and the method was proven to be more reliable and has been extensively used in differential expression analyses.

The ordinary *t*-test was computed by the scaled mean difference $\bar{X}_1 - \bar{X}_2/\sigma(X_1 - X_2)$, so there can be the cases when *t*-statistics are large accidentally due to small

within-group variance. Moderated t-statistic has corrected that through empirical Bayes method (function `eBayes` in `Limma`) by introducing the EB moderated variance estimators that takes into account the degree of freedom of the samples in the test. In this analysis, the results with FDR corrected p-value of less than `1e-4` were considered as significant.

3.3 Multiple data integration via joint latent variable model

The topic of integrating multiple data types has been of great interest in biology. Different biological information are usually collected over the same samples. Being able to integrate those data would be particularly useful in discovering sub-patterns of heterogeneous data such as discovering disease subtypes for patient stratification. In this thesis, a question to be addressed is whether the genetic profiles of transporters are able to recapitulate the cancer subtypes from the samples. As the data were acquired from multiple sources, a suitable statistical tool for integrating those data are of interest to the project.

In general, algorithms for data integration are based on matrix factorisation (MF) with various sparsity methods imposed. Early methods have limitations such as the number of datasets allowed in the model, or the datasets were learned separately. Klami *et al.* [25] developed a latent variable model aiming to explain the relationship between groups of variables called Group Factor Analysis (GFA) as an extension to the well-known factor analysis (FA) model. Flat gamma prior was used to impose sparsity and implement groupwise relationship. Kirk *et al.* [26] also developed a Bayesian joint latent variable model to cluster samples using multiple datasets using Dirichlet-multinomial allocation (DMA) mixture which can also accommodate time-series data. The method selected for this analysis is the widely used clustering method specifically implemented for biological data of different types such as gene expression (continuous) and mutation data (binary). It was developed by Shen *et al.* [12], and is available as an R package called `iCluster` [13]. The principles of the method are briefly explained below.

The general factorisation of a matrix $\mathbf{X}$ can be written as:

$$\mathbf{X} = \mathbf{WZ} + \epsilon \tag{1}$$

Given that $\mathbf{X}$ is the mean-centred matrix of genomic profiles of dimension $p \times n$, and $\mathbf{W}$ is the coefficient matrix where $\dim(\mathbf{W}) = p \times (K-1)$. $\epsilon = (\epsilon_1, \ldots, \epsilon_p)'$ is the error vector in which $\mu(\epsilon) = 0$ and $\mathbf{Cov}(\epsilon) = \mathbf{\Psi}$, whereby $\mathbf{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$.

$\mathbf{Z} = (\mathbf{z_1}, \ldots, \mathbf{z_{K-1}})'$ is the factor matrix. `iCluster` utilises the concept of the K-mean clustering algorithm, where the factor matrix Z is assigned to be indicator matrix of dimension $(K-1) \times n$ whereby each row represents the indicator vector of cluster k. The matrix is normalised so that they have a unit Euclidean norm, i.e. $||\mathbf{z_k}||_2 = 1$. That is

$$\mathbf{z_k'} = (\underbrace{0, \ldots, 0}_{\text{cluster 1:k-1}}, \underbrace{\frac{1}{\sqrt{n_k}}, \ldots, \frac{1}{\sqrt{n_k}}}_{\text{cluster k, length } n_k}, \underbrace{0, \ldots, 0}_{\text{cluster k+1:K}}) \tag{2}$$

n_k is the number of samples in cluster k, and $\sum_{k=1}^{K} n_k = n$. With this parameterisation, the objective function for K-means algorithm can be written, in normal form and using trace function, as:

$$\min \sum_{k=1}^{K} \sum_{C(i)=k} ||\mathbf{x_i} - \mathbf{m_k}||^2 \quad \equiv \quad \min(\text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{Z}')) \tag{3}$$

Note that with $\mu(\mathbf{X}) = 0$, $\text{tr}(\mathbf{X}'\mathbf{X}) = \text{Var}(\mathbf{X})$. Therefore, the equation above refers to variance of the data – variance between the clusters. Since $\text{Var}(\mathbf{X})$ is constant, this is equivalent to maximising the variance between groups:

$$\max_{\mathbf{Z}\mathbf{Z}'=\mathbf{I_K}} \text{tr}(\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{Z}') \tag{4}$$

Let us assume there are m types of input datasets, each data type denoted as $\mathbf{X_i}$ where $i = 1, \ldots, m$, and the factorisation form can therefore be written as $\mathbf{X_i} = \mathbf{W_i}\mathbf{Z} + \epsilon_\mathbf{i}$. Note that the factor matrix Z is jointly learnt for all the data types. In biological context, $\mathbf{Z}$ would refer to the molecular subtypes of the samples. On the other hand, ϵ is an independent error term and represents the unique variance of the data.

To be able to solve the equation above using maximum likelihood estimation (MLE) of Gaussian latent variable model, $\mathbf{Z}$ needs to be approximated into continuous values. This way would unfortunately reduce the interpretability in terms of K-means clustering, but would also reduce the complexity of finding the optimal $\mathbf{Z}$. Let $\mathbf{Z}^*$ be defined as the continuous parameterisation of $\mathbf{Z}$ while assuming that $\mathbf{Z}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$; with this assumption, the input matrices can be written in the form of multivariate Gaussian as follows:

$$\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_m})' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{5}$$

where $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}' + \mathbf{\Psi}$.

The method employs EM algorithm to compute the complete data log-likelihood:

$$\begin{aligned} \ell(\mathbf{W}, \mathbf{\Psi}) = &- \frac{n}{2} \left(\sum_{i=1}^{m} p_i \log(2\pi) + \log\det(\Psi) \right) \\ &- \frac{1}{2} \left(tr((\mathbf{X} - \mathbf{W}\mathbf{Z}^*)'\Psi^{-1}(\mathbf{X} - \mathbf{W}\mathbf{Z}^*)) + tr(Z^{*'}Z^*) \right) \end{aligned} \tag{6}$$

`iCluster` sets sparsity by adding Lasso regularisation term (see in section 3.4) to the log-likelihood function (Equation 6), controlled by parameter λ. In this analysis, the non-sparse solution ($\lambda = 0$) as well as the sparse solution ($\lambda = 0.01$) were compared, while a range of K values were tested to observe how samples are clustered together. The data used as the input are the expression of 416 transporter genes and their mutation profiles.

3.4 Feature selection: detection of significant gene - drug pairs

Biological data in general are heterogeneous and the distribution of expression of a gene across different samples is rather inconsistent, making them difficult to analyse with a particular method and to draw conclusions from their results. Multiple methods listed below were chosen to capture the relationships in different ways with different levels of complexity aiming to identify the strong relationships for gene-drug pair candidates.

Correlation test

The calculation of correlation coefficients is based on an assumption of data being normally distributed. However, the correlation test has been proven to be a simple yet powerful tool to identify the association among variables despite the non-uniform nature of data. The analyses in this study were done in both pan-cancer and tissue-specific manners. `cor.test` function in *R* was used to calculate Pearson and Spearman correlation coefficients for each gene-drug pair and their associated p-value.

For pan-cancer analysis, the correlation test was performed in all possible combinations of genes and drugs. The correlation coefficients along with p-value from correlation test were stored. The p-values were FDR corrected and those with p-value of less than 1×10^{-3} were identified as significant.

Tissue-specific analysis is relatively more difficult. The analysis suffers from both the amount of cell lines in each tissue and the poor sensitivity of the majority of cell lines in that tissue for some drugs. Therefore, a threshold was set for a drug to be used in the test for each tissue type as follows:

For a drug to be analysed in a particular cancer type, there must be

1. at least 8 cell lines being screened. Preliminary analysis found that the correlation test of fewer than 8 samples are not reliable.

2. those cell lines must constitute to at least 20% of all number of cell lines in the tissue type.

3. at least 3 cell lines having IC_{50} below the maximum concentration of the drug. This is biological filtration of the data. If the majority of the cell lines in the tissue are insensitive to a drug, then the drug-tissue pair was removed from the analysis to avoid false conclusion.

Analysis of Variance

The analysis of variance (ANOVA) of a gene-drug pair were performed to observe the significance of the drug response between the groups. The groups here are defined as samples with high and low expression levels. The main issue for using this method is that the gene expression levels were originally represented as continuous values, and two mixture models were unsuccessful to categorise the samples. An easy but robust way to perform this analysis is to consider cell lines with top 10% of expression

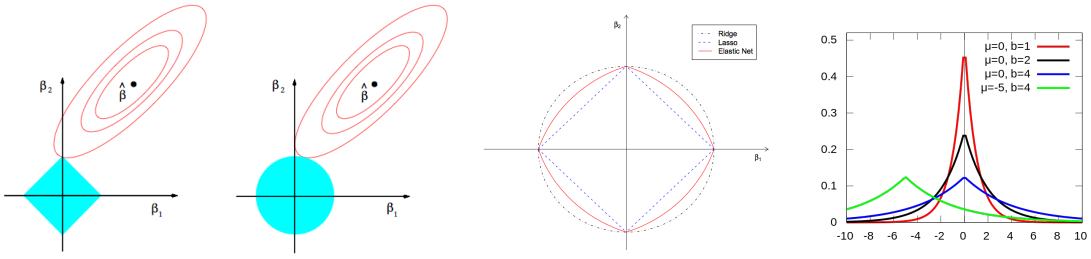
levels as 'highly expressed' and the bottom 10% as 'lowly expressed'. The samples in between those two ranges were neglected. Ordinary t-test was then performed to the binarised samples, with the tissue effect taken into account as a covariate (control variable) of the analysis. However, due to the 80% loss of samples during binarisation, an alternative method based on k-mean clustering was proposed in Chapter 5.

Linear model with regularisation

To apply linear models, the relationship between independent variables (drug sensitivity, AOC) and dependent variables (gene expression) in this analysis was assumed to be linear. The common problem here, which is similar to other data-driven biological analyses, is the high dimensional feature space ($p >> N$ problem). To avoid overfitting, linear model with regularisation was introduced to select most relevant features.

The two most renowned regularisation methods are Ridge and Lasso (least absolute shrinkage and selection operator) regularisation. Ridge uses ℓ^2 penalty term, limiting the size of the model coefficients while Lasso uses ℓ^1 penalty, imposing sparsity to the coefficient matrix and hence the model is more interpretable. From Bayesian point of view, Ridge regression is equivalent to the case when coefficients are assigned to normal prior distribution while Lasso has Laplace prior which are sharply peaked at their means.

The illustration comparing Ridge, Lasso, and elastic net regularisation is summarised in Figure 6:



(a) (b) (c)

Figure 6: Geometrical representation of different regularisation methods. Figure 6a shows two-dimensional space geometrical representation of Lasso (left) and Ridge (right) regularisation. The contour shows the regularisation path as the optimal $\hat{\beta}$ is regularised. Lasso regularisation applies ℓ^1 penalty, allowing the contour to reach the axis and imposing sparsity. The Ridge regression follows ℓ^2 penalty and therefore sparsity cannot be imposed. [Figure from Bishop [27]]. Figure 6b illustrates the geometrical representation of Lasso, Ridge, and Elastic Net, which is the combination of ℓ^1 and ℓ^2 penalty [28], and Figure 6c is the Bayesian representation of Lasso regularisation as having Laplace prior distribution.

Consider the standardised predictor matrix $\mathbf{X}$ of dimension $n \times p$ and $\mathbf{Y}$ as the

n-vector of observation. The Lasso regularisation follows:

$$\min_{\beta} ||\mathbf{Y} - \mathbf{X}\beta||^2 \quad \text{subject to} \quad ||\beta||_1 \leqslant t \tag{7}$$

Where t is a parameter determining the amount of regularisation and $||Z||_p = \left(\sum_{i=1}^{N} |Z_i|^p\right)^{1/p}$ is the ℓ^p norm.

The equation 7 can be rewritten by introducing Lagrange multiplier λ as:

$$\min_{\beta} \left\{ ||\mathbf{Y} - \mathbf{X}\beta||^2 + \lambda ||\beta||_1 \right\} \tag{8}$$

However, Lasso is greatly limited by the number of samples. In $p > n$ cases, Lasso only selects at most n variables. While this can be sufficient for the task, Elastic net regularisation was instead used for this analysis. Elastic net was developed by Zou & Hastie [28] by adding Ridge penalty term to Lasso. Elastic Net can usually outperform Lasso in $p > n$ cases and can select correlated predictors together, which were the advantage of Ridge regression.

The Elastic net modified the Equation 8 by adding Ridge penalty term, which is to solve:

$$\min_{\beta} \left\{ ||\mathbf{Y} - \mathbf{X}\beta||^2 + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right] \right\} \tag{9}$$

The Elastic net regression was implemented in R in the package `glmnet` [29]. The equation 9 was solved over a range of λ. The trade-off between Lasso and Ridge is controlled by α, where the default value is $\alpha = 1$, and therefore the model follows Lasso regularisation. This value can be adjusted to have a lower value (min=0) to add Ridge penalty term. In this analysis, 10-fold cross validation was applied. `glmnet` returns two values of λ, the `lambda.min` which refers to value of λ at the lowest cross validation error and `lambda.1se`, the λ in the range of 1 standard error from `lambda.min` but returned more regularised and therefore results in simpler model. In this analysis, the result from `lambda.1se` was compared with `lambda.min` and the results are discussed.

3.5 Learning the transporter interactions during drug intake via hierarchical group-Lasso

Two or more transporters may function together in drug uptake. Their individual effect may not be strong enough to capture in Lasso or Elastic net regularisation, so a method for interaction learning among predictors are therefore one focus of this thesis.

The topic of interaction learning is an active study field. Logic regression [30], for example, finds the interaction relationship by using boolean combinations of most predictive features, e.g. $(G_1 \wedge G_2) \vee G_3$, where G_i indicates arbitrary gene. This method is capable of discovering higher order (i.e. more than two features) interactions, but it is limited to binary variables and cannot directly be applied in this case.

Most of the interaction learning methods that can accommodate continuous variables where scalability is an issue are penalty-based. The method used in this analysis is hierarchical group-Lasso regularisation [14]. The method can model pairwise interactions of both categorical and continuous variables.

Biological data is generally acquired by collecting several attributes of samples whereas the number of samples are much lower ("$p >> n$" problem). By modelling interaction among the features would further increase the size of the feature space to $p + \binom{p}{2}$. For 416 SLC and ABC genes of interest, the total number of features in the model was over 100,000. The method proposed by Lim & Hastie [14] is called `GLINTERNET` which is based on group-lasso regularisation. With p in this order of magnitude, the model can directly be applied to all variables. However, larger scale of p may need variable screening to reduce the dimension first.

Let the feature space denoted as $\mathbf{X}$, and $\mathbf{X_{i:j}}$ represents the interaction vector $\mathbf{X_i} * \mathbf{X_j}$. The multiplication applies in element-wise manner for each row, i.e.

$$\mathbf{X_i} * \mathbf{X_j} = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{bmatrix} * \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} = \begin{bmatrix} x_{1i}x_{1j} \\ \vdots \\ x_{ni}x_{nj} \end{bmatrix} \tag{10}$$

Imposing hierarchy to first-order interaction model

Hierarchy for interaction model refers to the dependency of interaction effects to the presence of main effects. `GLINTERNET` was considered as a suitable method in this analysis partly because the model obeys strong hierarchy. Strong hierarchy is the case when an interaction is detected only if its main effects are present. To put it in biological context of this thesis, this makes sense to assume that two transporters that act together are those that individually affect the drug sensitivity, but their combination effect was unknown, or was not strong enough.

In general, the first-order interaction can be modelled by adding the additional interaction terms to the linear model:

$$\mathbf{Y} = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i<j} X_{i:j}\beta_{i:j} \tag{11}$$

The algorithm applies group-Lasso regularisation, the more general version of Lasso to distinguish main and interaction effects and to be able to impose hierarchy. Assume the variables are divided into groups. Let denote p as the number of groups of features that do not necessarily be of the same size, and $\mathbf{X_i}$ are the feature matrix of group i. The group-Lasso estimates the optimal coefficients β by modifying Lasso (equation 8) into:

$$\underset{\beta,\beta_0}{\operatorname{argmin}} \left\{ \frac{1}{2}||\mathbf{Y} - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{X_j}\beta_j||_2^2 + \lambda \sum_{j=1}^{p} \gamma_i||\beta_j||_2 \right\} \tag{12}$$

The optimal coefficients $\hat{\beta}_j$ can be obtained, and similar to the previous cases, λ controls the regularisation. The term γ were added to adjust the group penalty. In

this case, $\gamma_j = 1 \ \ \forall j$ for simplicity. `GLINTERNET` generates a range of λ to apply to the model, starting from a very large one to regularise most of the variables away. It is often observed that the choice of λ selected from this algorithm can be too large that several sets of λ values needed to be generated. The optimal λ values are found through cross validation. In this study, 10-fold cross-validation was used.

Overlapped group-Lasso

Overlapped group-Lasso is group-Lasso that allows a variable to be in different groups. The overlapped group penalty imposing strong hierarchy is added to equation 12. `GLINTERNET` has been formulated by using overlapped group-Lasso so that it could accommodate multi-level categorical data. In this analysis, the variables (gene expression) are continuous.

The generalisation of Equation 10 is given by:

$$\begin{aligned} \mathbf{X_{1:2}} &= [1 \ \ \mathbf{X_1}] * [1 \ \ \mathbf{X_2}] \\ &= [1 \ \ \mathbf{X_1} \ \ \mathbf{X_2} \ \ (\mathbf{X_1} * \mathbf{X_2})] \end{aligned} \tag{13}$$

The linear model for the interaction of continuous variables is the extension of Equation 8 and 12, which follows:

$$\underset{\beta}{\operatorname{argmin}} \frac{1}{2} ||\mathbf{Y} - \beta_0 \cdot \mathbf{1} - \mathbf{X_1}\beta_1 - \mathbf{X_2}\beta_2 - ([1 \ \mathbf{X_1}] * [1 \ \mathbf{X_2}])\beta_{1:2}||_2^2 + \lambda(||\beta_1||_2 + ||\beta_2||_2 + ||\beta_{1:2}||_2) \tag{14}$$

For this analysis, `GLINTERNET` was performed to each response vector (sensitivity for one drug) without variable screening. The results are discussed in the next chapter.

4 Results

4.1 The Landscape of Transporter Expression

Transporter genes are expressed differently in different tissues. The information of which tissues a specific transporter is mostly expressed have been well studied and collected in www.bioparadigms.org. However, the landscape of transporter expression in cancer, which can be different from normal tissues, has not been well investigated and this could further connect to tumorigenesis and consideration for choices of treatment. For instance, knowing which transporters are excessively expressed in the tissue of interest can lead to appropriate selection of drugs that found to be specific to the transporter.

4.1.1 Tissue-specific patterns of SLC and ABC transporters

To get a glimpse of how transporters are differently expressed in different tissues, the heatmap showing average expression level of all the cell lines in each tissue are shown for SLC and ABC genes in Figure 7.

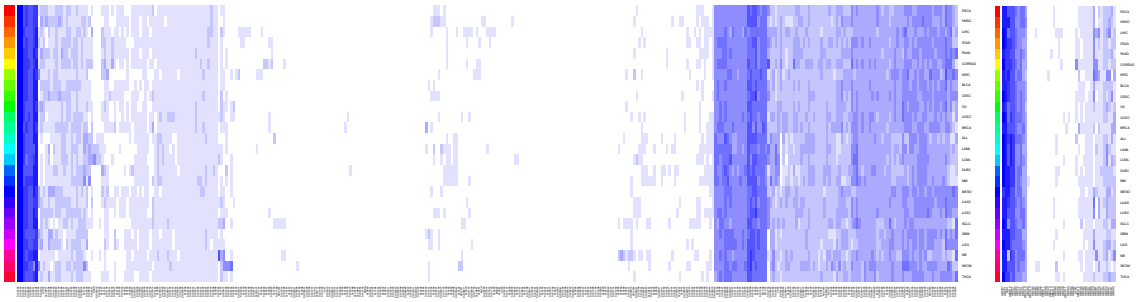
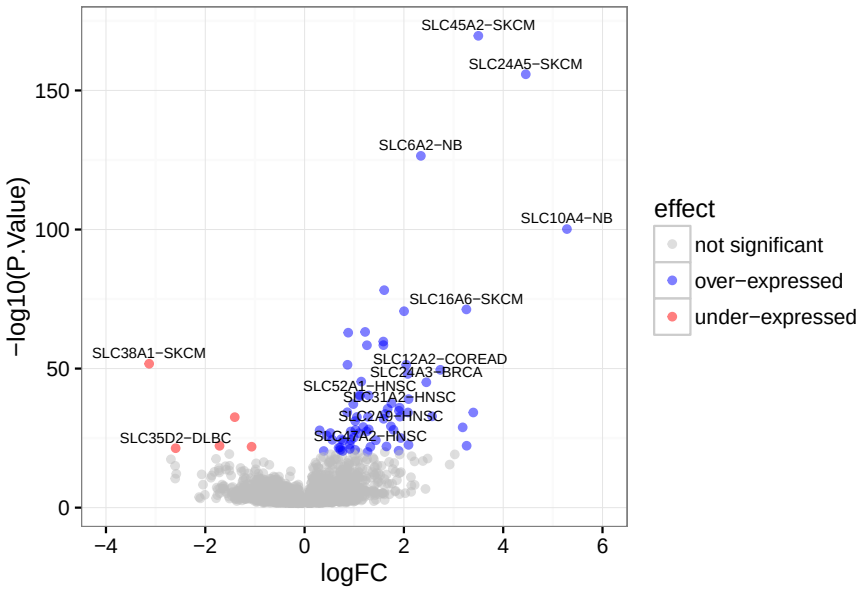


Figure 7: The heatmap of SLC expression and ABC expression level in different tissues. Each cell represent the average expression level of the gene (row) and cancer tissue (column). The SLC expression profile shown in left panel while ABC are shown in right panel. The rainbow on the left of each panel represents the tissue colour code.

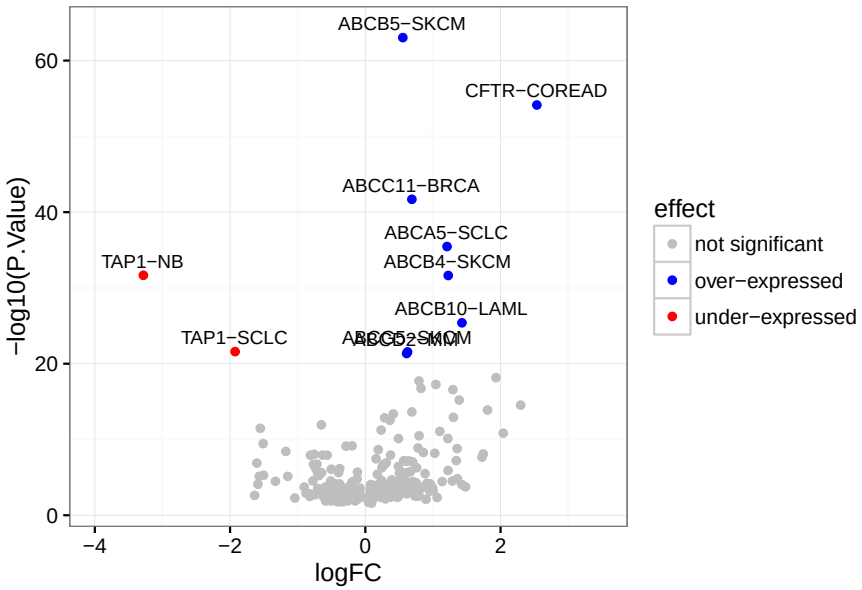
The column of the data is clustered using Euclidean distance. The leftmost columns of both panels show transporters that are widely expressed in the majority of samples, thus they exhibit 'house keeping' properties. Those cases are particularly interesting since under-expression of genes that are normally highly-expressed might be associated with cancer progression.

The identification of tissue-specific transporters were calculated by moderated t-test with FDR corrected p-values through package `Limma` in R (details in Section 3.4). Since the algorithm provides several candidates with low p-values, the threshold for significance is p-value $<$ `1e-20`.

In Figure 8, the results from differential expression analysis of SLC and ABC genes were shown. The corresponding table of this analysis was included in Appendix A.



(a) SLC differential expression analysis



(b) ABC differential expression analysis

Figure 8: The scatter plot of differential expression analysis for each tissue type of SLC and ABC genes. Each point represents a gene-tissue pair and the plot shows -log(p-value) against the log fold change of the expression for each gene-tissue pair.

There are notable cases, especially those conferring under expression, which are genes that are normally highly expressed. An example is the under expression of TAP1 gene in neuroblastoma (NB) and small-cell lung cancer (SCLC), as shown in the left side of Figure 8b. TAP1 is the **t**ransporter associated with **a**ntigen **p**rocessing; and it is required for the histocompatibility complex (MHC) class I pathway. MHC I is important for tumour surveillance, and the lack of this pathway has been linked to

tumorigenesis in neuroblastoma [31] [32] [33]. Similarly, lack of TAP1 in lung cancer (SCLC) have led to the same effect [34]. This further indicates the endogenous roles of transporters in cancer as suggested in Nigam [15].

While low expression of particular genes in ABC family are well documented, literature supports for similar cases in SLC genes were still largely insufficient. The analysis found underexpression of SLC38A1 in SKCM, SLC10A3 and SLC25A43 in SCLC, SLC16A1 in BRCA, and SLC35D2 in DLBC. Not many strong evidences to support the roles of these transporters in cancer progression were found.

In the opposite cases, there are significantly more tissue-specific genes exhibiting over expression. The strongest case found is SLC45A2 in skin cancer (SKCM), which has been supported by recent literatures about SLC45A2 involvement in pigmentation pathway is a strong risk factor in melanoma [35]. Other recent studies also confirmed the importance of transporter SLC24A5 in controlling pigmentation in different ethnicities [36] [37]. Over expression of this gene is therefore likely to induce effects that lead to melanoma.

There are other several cases that differential expression have been captured in this analysis and their association to tumour progression have been known, while there are also many others, which their effect might not be as strong, but this could provide further novel insights into more detailed studies of these transporters and their roles in tumorigenesis.

4.1.2 Recapitulation of cancer subtypes characteristics based on multiple genetic profiles

Two data types, i.e. (1) standardised gene expression profiles of 416 SLC and ABC genes and (2) cancer functional events (hypermethylation and copy number variations) in binarised format were used as the input to compute the common factor loading $\mathbf{Z}$ using Equation 6 via `iCluster` in *R*.

The Figure 9 below shows the results of cancer cell lines clustering when assigning $k = 30$ and the algorithm performed in 50 iterations. The row represents true cancer subtypes, while the column represents 30 clusters from the algorithm. This is to observe how well transporter genetic profiles are able to recapitulate cancer heterogeneity, and which cancer tissues have distinct transporter expression patterns than the others.

From this analysis, it can be seen that cell lines in breast, skin, and large intestine show relatively strong transporter tissue specificity pattern, with more than 50% of all cancer cell lines in the tissues were assigned to the same cluster. This can lead to a hypothesis that SLC and ABC transporter genetic profiles in these tissues are able to recapitulate the cancer tissue characteristics more than other tissues. This can lead to a hypothesis of whether transporters in these tissues play more roles in tumorigenesis than the rest of the tissues.

The concept of matrix factorisation can further be applied by assigning $k = p$, and a cell line will belong to one cluster. Weight matrix in that scenario can be used for further prediction instead of gene expression profile themselves.

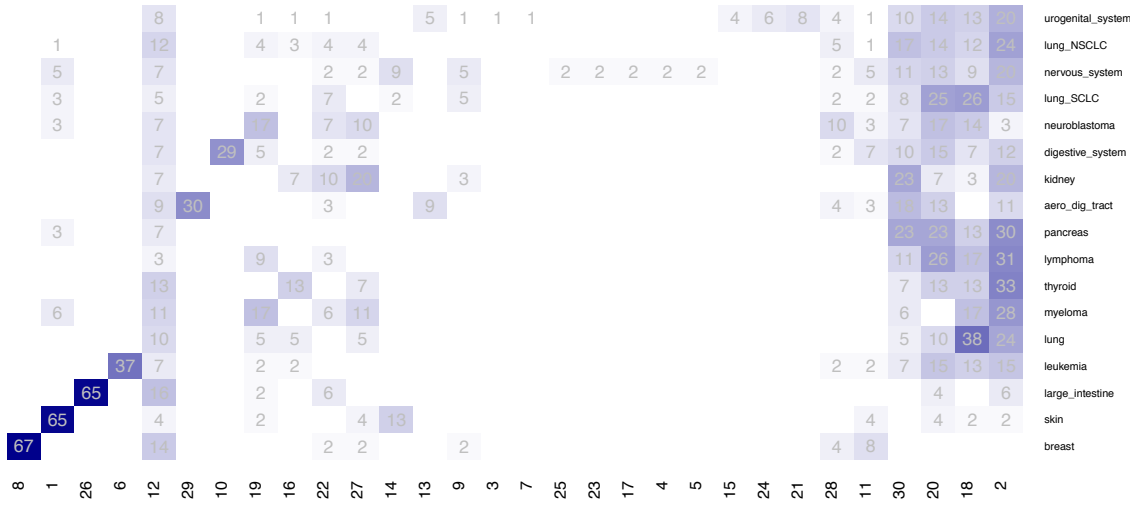


Figure 9: Cluster results from `iCluster` using $k = 30$ with 50 iterations. The number in each cell represents the percentage of cell lines for each cancer tissue (row) belonging to a cluster (column). Three cancer tissues i.e. breast, skin and large intestine show strongest within-tissue clusters.

4.2 The Transporter - drug association

4.2.1 Correlation analysis of drug-transporter association

Despite both of the independent (gene expression) and dependent (drug response) data are not normally distributed, simple correlation analysis has been proven useful in identifying significantly correlated patterns. This correlation analysis was performed in both pan-cancer and tissue-specific manner.

Pan-cancer correlation analysis

This Pearson's correlation test for each drug was performed to all genes. The correlation is identified as significant if the drugwise FDR corrected p-value of the test is smaller than `1e-3`. Drugs with SLC or ABC genes identified as significant were shown in the boxplot below in Figure 10, and the corresponding information from the correlation test for those labelled genes in Figure 10 is shown in Table 2.

Tissue-specific correlation analysis

Tissue-specific analysis has to encounter the issue of insufficient sample size. With the criteria described in Chapter 3, several drugs have been discarded from specific tissues. The heatmap showing availability of data for each tissue is shown in Figure 11.

There are 46 drugs that have been discarded from the tissue-specific analysis due to insufficient sample size. The names of those drugs are listed in Table 3.

The correlation test was performed to gene-drug pairs in all available tissues that met the criteria discussed in Section 3.4, resulting in over 1 million correlation outcomes generated. The FDR corrected p-values were performed for each drug, and the significant pairs are those with p-value lower than `1e-3`.

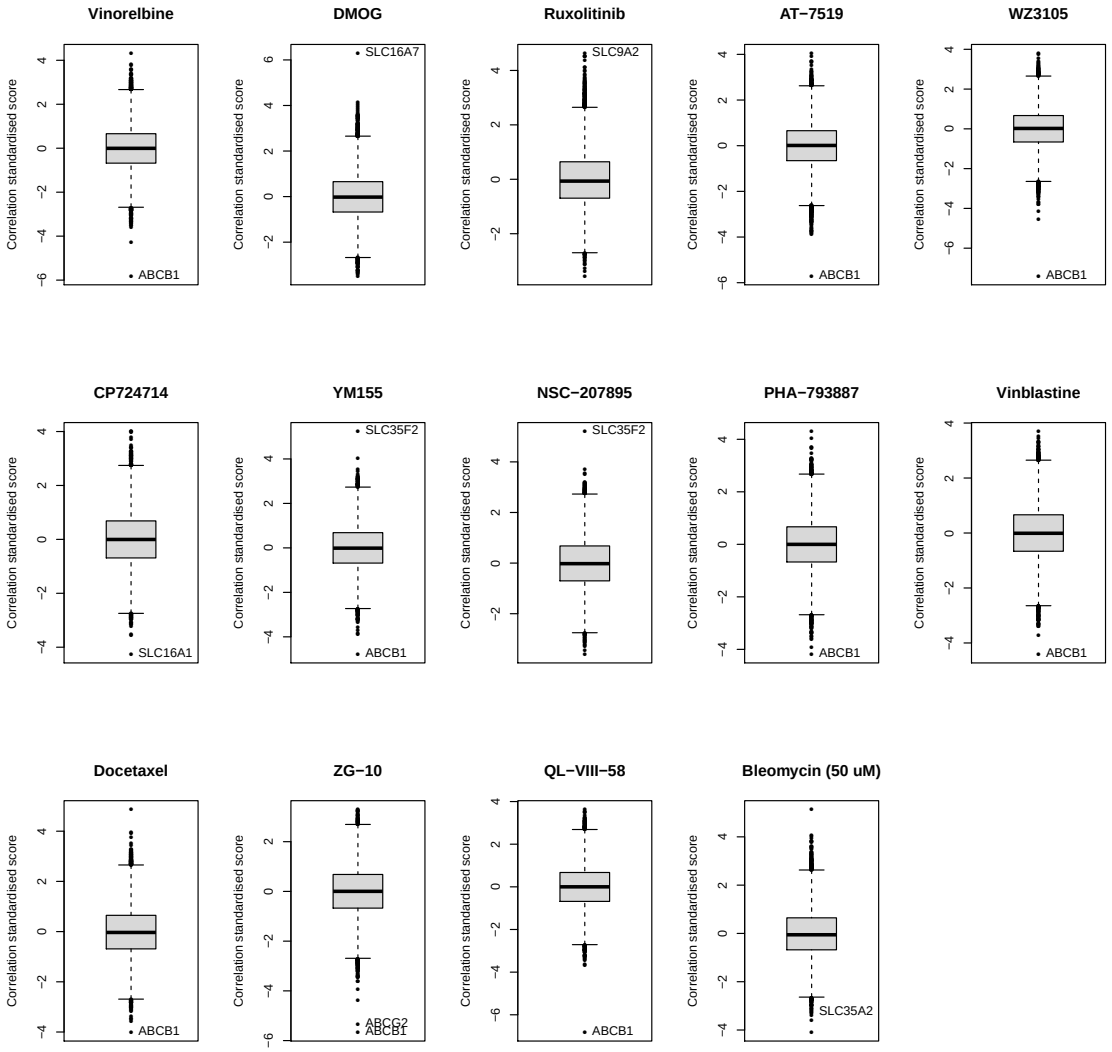


Figure 10: Boxplot shows standardised (z) values for correlation coefficients of all 17,419 genes for each drug. Out of 265 drugs. 14 drugs with SLC or ABC genes identified as significant are shown.

The significant associations were listed in the Table 4, sorted by tissue names. It was found that most significant associations identified have positive correlation coefficients, therefore conferring drug sensitivity.

Text mining was attempted to retrieve number of publications in PubMed containing the tissue, drug, and gene names in the list. However, there are variants of the names and many drugs are in investigational phases. The retrieval of number of publications as a support of the findings are therefore not trivial.

Note that the tissue-specific analysis was not applied in other association detection methods due to the sample size issues.

Table 2: Corresponding gene-drug associations for Figure 10.

Gene Name	Drug Name	z-value	FDR	Effect
ABCB1	Vinorelbine	-5.808	1.2E-06	resistant
SLC16A7	DMOG	6.288	1.4E-06	sensitive
SLC9A2	Ruxolitinib	4.628	6.5E-04	sensitive
ABCB1	AT-7519	-5.702	3.3E-05	resistant
ABCB1	WZ3105	-7.394	8.7E-11	resistant
SLC16A1	CP724714	-4.253	2.0E-04	resistant
ABCB1	YM155	-4.773	2.6E-04	resistant
SLC35F2	YM155	5.236	6.7E-05	sensitive
SLC35F2	NSC-207895	5.202	3.3E-05	sensitive
ABCB1	Vinblastine	-4.402	1.8E-04	resistant
ABCB1	ZG-10	-5.660	6.5E-06	resistant
ABCG2	ZG-10	-5.338	3.5E-05	resistant
ABCB1	QL-VIII-58	-6.815	1.4E-12	resistant
SLC35A2	Bleomycin (50 uM)	-3.268	6.6E-04	resistant

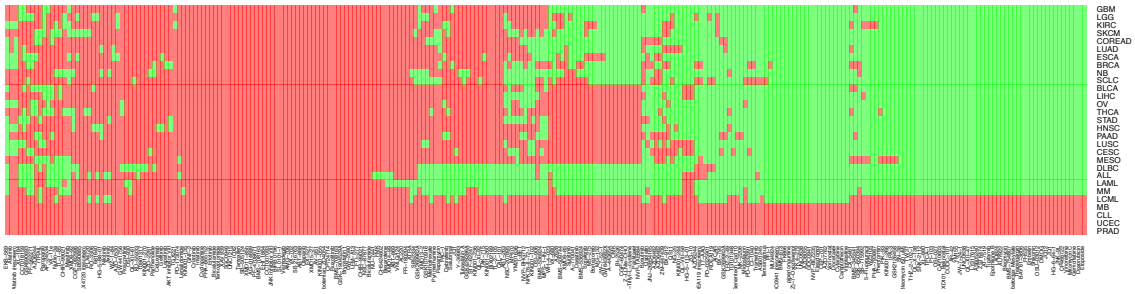


Figure 11: Tissue-specific analysis suffers from insufficient sample size. Tissue-drug pairs that the correlation tests were not performed are shown in red. A number of drugs were entirely discarded from the analysis.

Table 3: Drugs that were removed from the tissue-specific analysis due to insufficient sample size.

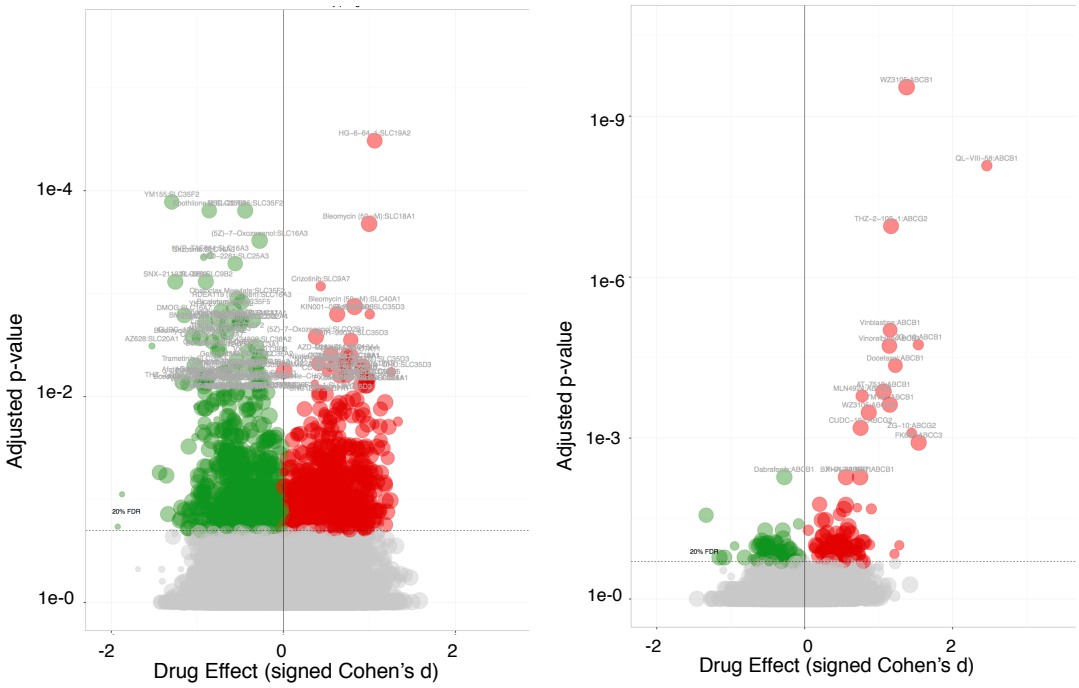
NSC-87877	Bicalutamide	CHIR-99021	FTI-277	LFM-A13	GW-2580
BMS-708163	Ruxolitinib	CP724714	Zibotentan	KIN001-055	AV-951
Olaparib	ABT-888	SB 216763	VX-702	AMG-706	Vismodegib
JNK Inhibitor VIII	CCT007093	EHT 1864	BMS-708163	PF-4708671	AG-014699
Tamoxifen	IOX2	UNC1215	SGC0946	XAV 939	Temozolomide
PHA-665752	Cyclopamine	Imatinib	Crizotinib	GNF-2	KIN001-135
Bryostatin 1	GSK-1904529A	XMD15-27	EX-527	SL 0101-1	BIRB 0796
XMD11-85h	SB-505124	Bicalutamide	Erlotinib		

Table 4: The result from correlation test for tissue-specific correlation analysis.

Tissue	Drug name	Gene	ρ	df	adj. p-val	z	Effect
ALL	PLX4720	SLC13A4	0.858	21	9.2E-05	3.255	sensitive
ALL	PLX4720	SLC26A3	0.891	21	2.0E-05	3.382	sensitive
ALL	PLX4720	SLC28A3	0.877	21	3.2E-05	3.329	sensitive
ALL	PLX4720	SLC9A5	0.809	21	9.8E-04	3.068	sensitive
ALL	PLX4720	ABCA13	0.851	21	1.1E-04	3.227	sensitive
BRCA	RDEA119	SLCO3A1	0.842	41	1.4E-08	3.428	sensitive
BRCA	BMN-673	SLC7A7	0.784	43	2.1E-06	3.458	sensitive
BRCA	BMN-673	SLC51B	0.706	43	2.1E-04	3.120	sensitive
BRCA	BMN-673	SLC12A3	0.764	43	5.4E-06	3.373	sensitive
BRCA	Trametinib	SLCO3A1	0.697	43	7.4E-04	2.980	sensitive
BRCA	AS605240	SLC18A1	0.679	45	5.2E-04	3.253	sensitive
BRCA	RDEA119	ABCB1	0.799	41	6.4E-07	3.251	sensitive
BRCA	PD-0325901	ABCB1	0.722	40	3.1E-04	2.995	sensitive
BRCA	Trametinib	ABCB1	0.690	43	7.4E-04	2.952	sensitive
BRCA	AS605240	ABCB1	0.774	45	1.8E-06	3.701	sensitive
COREAD	Temsirolimus	SLC6A15	0.723	39	9.2E-04	3.196	sensitive
COREAD	GSK1070916	SLCO1C1	0.687	42	9.5E-04	3.061	sensitive
COREAD	Sunitinib	SLC4A4	0.939	13	9.5E-04	3.472	sensitive
COREAD	YM155	ABCB1	-0.738	43	7.5E-05	-3.290	*resistant*
DLBC	KIN001-270	SLC52A3	0.788	25	1.1E-04	2.857	sensitive
DLBC	KIN001-270	SLC26A3	0.862	25	3.3E-06	3.145	sensitive
DLBC	KIN001-270	SLC44A4	0.802	25	6.7E-05	2.913	sensitive
DLBC	KIN001-270	SLC44A3	0.766	25	2.2E-04	2.772	sensitive
DLBC	KIN001-270	SLC9A2	0.774	25	1.8E-04	2.802	sensitive
DLBC	KIN001-270	SLC47A1	0.842	25	7.8E-06	3.067	sensitive
DLBC	NVP-TAE684	SLC26A4	0.841	22	7.8E-04	3.166	sensitive
DLBC	BMS-536924	SLC51B	0.913	21	6.2E-06	3.368	sensitive
DLBC	KIN001-270	ABCC3	0.737	25	7.0E-04	2.656	sensitive
DLBC	NVP-TAE684	ABCC9	0.841	22	7.8E-04	3.166	sensitive
LAML	ATRA	SLC47A1	0.868	22	6.4E-05	3.314	sensitive
LAML	Dasatinib	SLC4A8	0.844	24	3.0E-04	3.022	sensitive
LAML	A-770041	SLC44A5	0.916	24	2.6E-07	3.141	sensitive
LAML	A-770041	SLC35F4	0.903	24	6.9E-07	3.097	sensitive
LAML	A-770041	SLC4A8	0.806	24	8.6E-04	2.769	sensitive
LAML	A-770041	SLC4A4	0.832	24	2.3E-04	2.859	sensitive
LCML	Nilotinib	SLC10A2	-0.974	8	7.7E-04	-3.296	*resistant*
LCML	Nilotinib	ABCA6	-0.981	8	5.0E-04	-3.318	*resistant*
NB	RDEA119	ABCC3	0.843	26	6.0E-05	3.428	sensitive
NB	CI-1040	ABCC3	0.906	25	7.6E-07	3.743	sensitive
NB	PD-0325901	ABCC3	0.851	26	9.0E-05	3.532	sensitive
NB	AS605240	ABCC3	0.862	25	3.7E-05	4.121	sensitive
NB	LY317615	ABCC3	0.842	25	3.7E-04	4.061	sensitive
OV	ABT-263	SLC12A3	0.797	30	2.8E-04	3.344	sensitive
OV	Afatinib (rescreen)	SLC23A2	0.787	30	3.4E-04	3.607	sensitive
OV	GSK-650394	ABCD2	0.912	16	7.1E-04	4.084	sensitive
SCLC	GSK-650394	ABCA10	0.656	49	7.1E-04	2.915	sensitive
SKCM	MP470	SLC16A8	0.812	48	3.9E-09	3.422	sensitive
STAD	AZD6244	SLC2A2	0.910	18	2.1E-04	4.090	sensitive
STAD	MP470	ABCC9	0.967	18	8.6E-09	4.099	sensitive

4.2.2 The analysis of variance (ANOVA)

The ANOVA results for ABC and SLC genes are shown in the figure 12.



(a) SLC-drug association (b) ABC-drug association

Figure 12: The ANOVA results, separately analysed for SLC and ABC genes. Each dot represents the gene-drug association pair, plotted in FDR adjusted p-value against Cohan's effect size. Associations marked in red are the pairs that confer resistance while the ones in green are those that confer sensitivity. The grey dots represent association below the significant threshold (20% FDR). [Figure courtesy of Mi Yang]

The ANOVA results from Figure 12 shows that ABC genes are more strongly associated with drug resistance, with ABCB1 and ABCG2 appear to be most significant to several drugs.

In contrast, SLC genes do not exhibit strong effect towards drug resistance or sensitivity, with the number of association pairs in both cases are relatively equal. This result contradicted earlier assumption by previous studies of SLCs that they majorly act as influx transporters, i.e. they mediate drugs to enter the cell.

The list of top hits identified in Figure 12 were summarised in Table 5 below with drugs having similar effects to the same genes represented together.

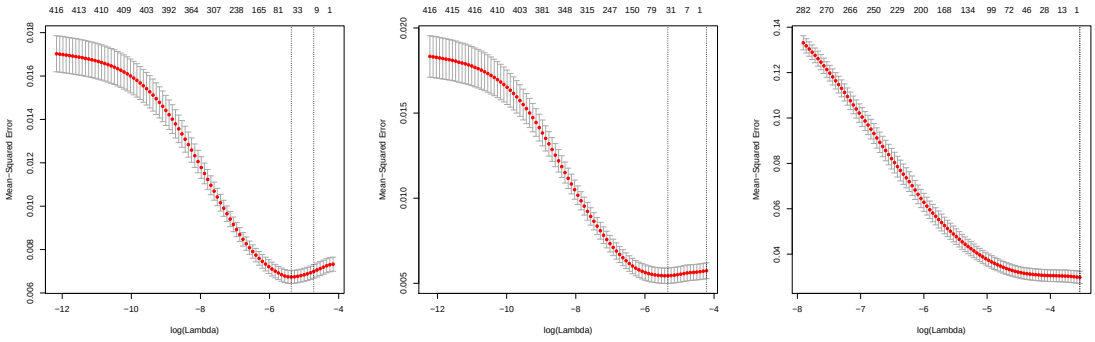
Table 5: Top hits from ANOVA, categorised by the drug that each gene confer sensitivity and resistance

Gene	Top sensitives	Top resistants
SLC20A1	AZ628, Trametinib	–
SLC9B2	SNX-2112, THZ-2-102-1	–
SLC16A7	DMOG	–
SLC25A24	Docetaxel	–
SLC6A18	CUDC-101	–
SLC35F2	YM155	–
SLC35D3	–	Z-LLNle-CHO, Bortezomib
SLC19A2	–	HG-6-64-1
SLC38A5	–	TGX221
SLC18A1	–	Bleomycin , Docetaxel
SLC25A1	–	UNC0638, MPS-1-IN-1, PXD101, Belinostat

4.2.3 Feature rankings with Elastic Net regularisation

The question to be addressed from this section is, which transporters, if any, are most associated with the drug response profiles. In addition to the previous methods, linear regression with regularisation was applied with the feature space containing 416 SLC and ABC transporter genes.

One of the issues observed is that the optimal regularisation parameter (λ) is not found. The plot of cross validation error against λ values are shown below for three drugs to show all possible behaviours of the models.



(a) Piperlongumine (b) Nutlin-3a (c) VX-680

Figure 13: Mean squared error (MSE) value against λ plot for three drugs. As a higher value of λ applied, the model is more regularised and the coefficients of many features are set to zero. The λ values that can be returned from the algorithm is the λ_{min}, which is the value when mimimum MSE is reached (vertical dotted line on the left of each plot), and the λ_{1se}, the more regularised value, shown in vertical dotted line on the right.

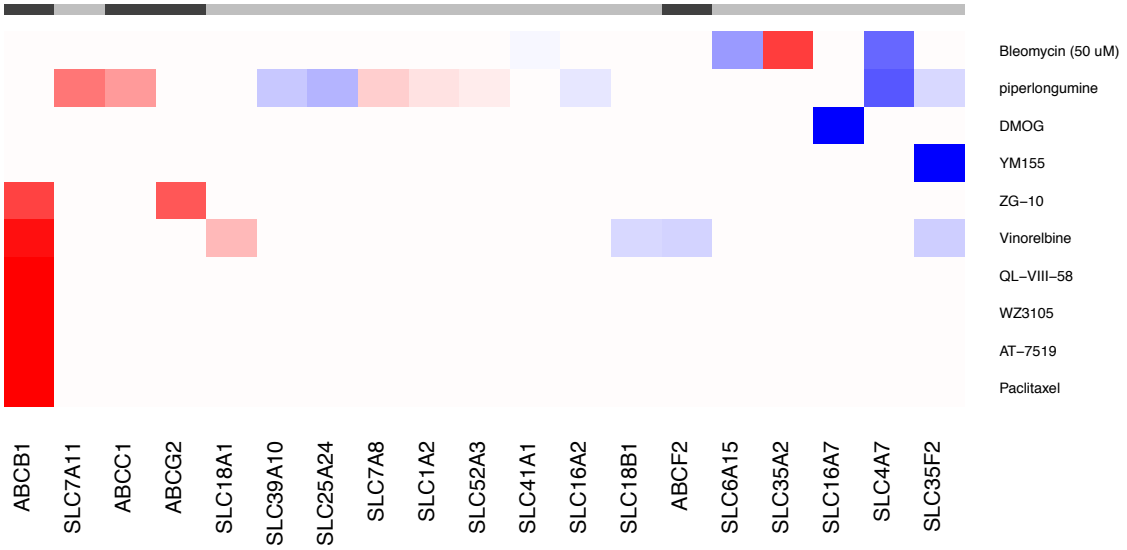
In Figure 13(a), the value of λ_{min} was found, with the λ_{1se} does not regularise all

features while Figure 13(b) is the case when λ_{1se} regularise all features, indicating that the optimality of λ_{min} is not significant. Figure 13(c) is the scenario when λ_{min} is not found, and no features were reported from the model.

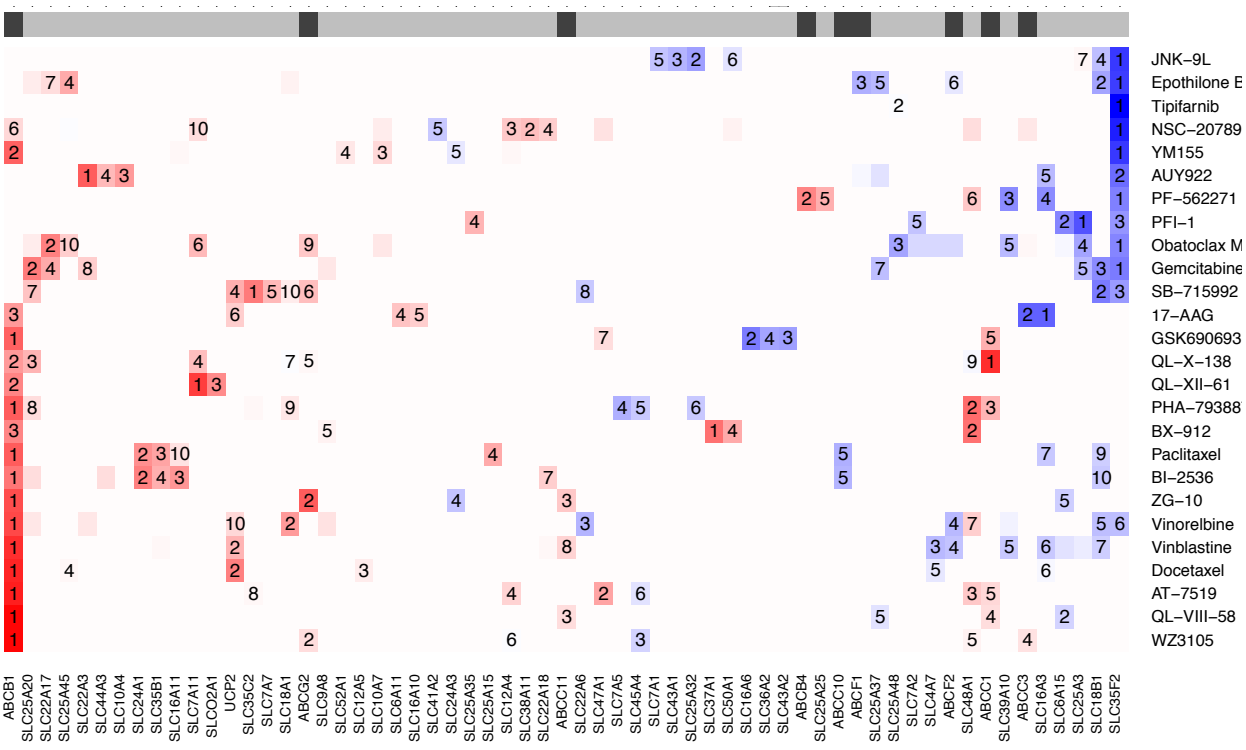
Using λ_{1se} as the regularisation parameter for all models, there are only 10 drugs (out of 265) that the model returns associated features. On the contrary, models with λ_{min} as regularisation parameters are less regularised and up to 161 drugs have associated features. The results of both models are shown in Figure 14a and Figure 14b. With the higher stringency of the model with λ_{1se}, the results can be considered as more reliable.

The analysis in Figure 14a, despite not showing many drugs, it supports known functions of ABCB1 as a multidrug resistance transporter, with it appeared to be the only transporters selected from the model in several drugs. Similarly, the results shown in Figure 14b shows similar effects, with ABCB1 ranked as the most associated transporters to several drugs. However, new information obtained from this analysis is the sensitivity of several drugs to SLC35F2 (shown in the rightmost column). It was known to be associated with YM155, which was also captured from the model, but it has never been known before that SLC35F2 were found sensitive in many other drugs (it ranked top 3 in 11 drugs). This can lead to further hypothesis of 'multi-drug sensitivity transporter' of SLC35F2, in the contrary to the property observed in ABCB1. Further analyses of the function of the transporter and the structures of its associated drugs will be needed to investigate and confirm this observation.

Figure 14b also represents several transporters associated to some drugs. This can be hypothesised either as (1) it suggests the redundancy of transporter functions, with a drug do not necessarily be associated to a particular transporter, or (2) they function cooperatively. To capture the dependency of two transporters for drug intake, the identification of transporter interactions will be discussed in the next section.



(a) The results obtained from Elastic Net feature selection, and λ_{1se} was chosen. The models were regularised that most drugs do not yield any results. On the other hand, the top hits represented by this model is in a higher confidence level than the results using λ_{min} in Figure 14b. The colour in each cell represents the weight of the transporter in the model, where blue confer sensitivity and red confer resistance. The column side colour on the top of the plot indicates whether the feature belongs to SLC or ABC superfamily.



(b) The similar representation to Figure 14a, but showing the results for models with λ_{min}. The models less regularised, resulting in more number of association in each drug. The plot shows results for a subset of drugs that ABCB1 or SLC35F2 were ranked in top 10.

Figure 14: Elastic Net feature selection results from λ_{1se} and λ_{min} as regularisation parameters respectively.

4.3 Interaction of multiple transporters in drug response

This section aims to disclose the information of two transporters that cooperatively mediate drug sensitivity or resistance, i.e. their effects are dependent on each other. This can be the case when, for instance, a drug needs to reach its target in a membraned organelle (e.g. nucleus, vesicle or ER), so it needs two transporters that associate with a drug (one at the cellular membrane and the other is at the organelle membrane), or when two transporter genes encode for proteins that function together as a complex. This association of transporter to drug sensitivity will from this point be referred as 'interaction'.

The first-order interaction among transporters were identified via `GLINTERNET` with default setting of the package. The algorithm reports main effects as well as interaction effects for each drug. Among 265 drugs, there are 109 drugs that interaction between two transporters were captured. The graph representation of all interaction detected from the algorithm of all drugs is shown in Figure 15.

gene1	gene2	n drugs
SLC6A11	SLC6A12	10
ABCC8	ABCC9	9
SLC5A12	SLC7A9	8
SLC20A1	SLC26A7	7
SLC1A3	SLC3A1	7
SLC20A2	SLC41A2	7
SLC25A20	SLC37A2	6
SLC6A6	SLC7A11	6
SLC41A2	SLC9A5	6
SLC39A9	UCP2	6
SLC12A9	SLC16A2	5
ABCA10	SLC1A3	5
SLC17A2	SLC1A7	5
SLC22A23	SLC2A9	5
MTCH1	SLC37A2	5
SLC25A45	SLC39A11	5
SLC39A14	SLC4A11	5
SLC34A1	SLC6A16	5
SLC26A5	SLCO1B1	5
ABCA10	ABCC9	4

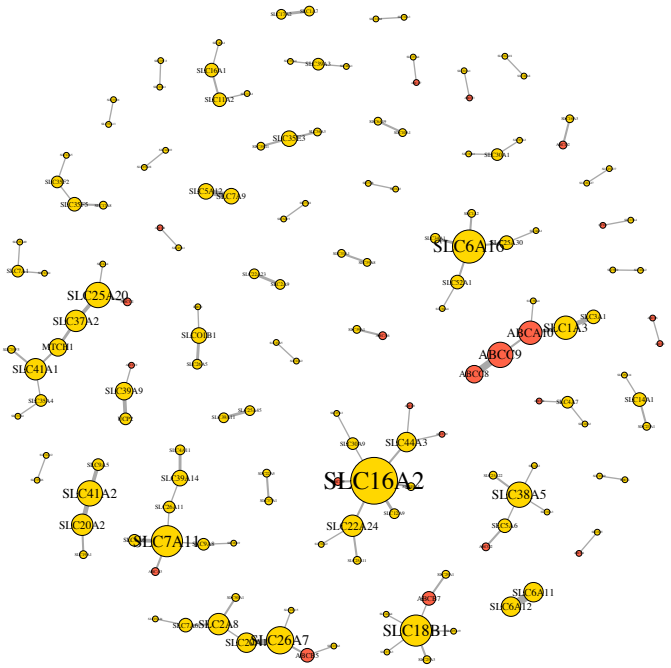


Figure 15: The landscape of transporter interactions from `GLINTERNET` with default hyperparameter settings, represented in graph.

In Figure 15, nodes represent genes with their sizes correspond to the number of interacting pairs. The colour of each node represents whether the gene belong to SLC (yellow) or ABC (orange) superfamily. Two genes are connected by an edge if the interaction between them is detected, and the width of an edge corresponds to the number of drugs the interaction is found. The table on the left of the figure shows top 20 transporter interaction pairs, sorted by the number of drugs that the interaction is detected.

Some interacting pairs are transporters in the same family, e.g. SLC6A11 and

SLC6A12 found to be associated to 10 drugs, and ABCC8 is associated with ABCC9 in 9 drugs. These cases may be resulted from the complex formation of the transporters that transport function required the presence of both genes. On the contrary, there are notably high number of interacting pairs from transporter genes in different families. Many of them are isolated pairs (two nodes connected with an edge in Figure 15) while many can be represented as connected graph. SLC7A11 confers resistance to several drugs from the association analyses in the previous section, it is found to be highly associated with SLC6A6 which had not been detected in other previous sections. SLC7A11 also found to be a hub of a cluster, similar to SLC16A2, SLC6A16, and SLC18B1. These interaction information can be useful for additional information to concern in drug discovery to improve the efficacy of a drug.

5 Discussion and Outlook

5.1 Further landscape analyses to disclose transporter heterogeneity

In Chapter 4, the analyses in the matrix factorisation part were focused onto tumour heterogeneity given transporter information. The clustering can be done in the opposite way, i.e. using multiple datasets of the same set of transporter genes. With this formalisation, latent variable matrix for the combination of gene expression databases, e.g. GDSC and CCLE can be done. Shared information among the multiple data types will be captured and this might improve the accuracy for further analysis with drug efficacy association.

The tissue specificity identified in the previous chapter provide the information obtained from cell lines. To be able to draw a conclusion in clinical context, the comparison of transporter expression in cell lines (GDSC) need be compared with the primary tumour databases (samples extracted from patients from The Cancer Genome Atlas, TCGA). By doing so, the agreements between cell lines and patient transporter landscape information can be observed, and this can pinpoint whether effects observed in cell lines are likely to be the same in patient levels.

5.2 Gene-drug association top hits agreements among different methods

To systematically confirm the gene-drug association, experimental validation is required to ensure the effect of a transporter to drug sensitivity.

With various limitations and sensitivities of different association detection methods used in this project, the results obtained from each method can be very different. The list of candidates for experimental validation should therefore be the significant top hits that appear across multiple methods.

The list of top hits that were identified as significant in more than one method is shown below.

Table 6: Significant top-hits from multiple methods and their novelty level. In the Effect column, S and R are abbreviated from 'Sensitive' and 'Resistant', respectively. The methods included here are Pearson's correlation test (CT1), Spearman's correlation test (CT2), Analysis of Variance (ANOVA), Elastic Net model using all genes as features (EN1), and Elastic Net model using SLC and ABC genes as features (EN2)

Gene	**Drug**	**CT1**	**CT2**	**ANOVA**	**EN1**	**EN2**	**Effect**	**Novelty**
SLC16A7	DMOG	✓	✓	✓	✓	✓	S	Novel
SLC35F2	YM155	✓	✓	✓	✓	✓	S	Studied in [18]
SLC35F2	NSC-207895	✓			✓	✓	S	Novel
SLC7A11	(5Z)-7-Oxozeaenol				✓	✓	R	Novel
SLC7A11	piperlongumine				✓	✓	R	Mentioned [38]
SLC9A2	Ruxolitinib	✓				✓	S	Novel
SLC16A1	CP724714	✓				✓	R	Novel

From the Table 6, there are only two hop hits which were in agreement with all methods, which are SLC35F2-YM155 and SLC16A7-DMOG. The first association pair has thoroughly been studied in Winter *et al.* [18] and it was found that the sensitivity of YM155 reduced significantly when SLC35F2 was knocked out. However, similar analyses for SLC16A7 and DMOG has not been found.

Furthermore,it was found that many other drugs has reported high sensitivity with SLC35F2 (as represented in Figure 14b). The most notable case here is NSC-207895, which the association can be captured in three methods and this association has also not been known.

SLC7A11 is also another SLC transporter that confer resistance with several drugs, where most of their associations were captured in Elastic Net model, and therefore could be a potential candidate to detailed studies of the multiple drug resistance functions of transporters in SLC superfamily.

5.3 Interaction learning

The interaction terms may be modified as follows.

By the formalisation of interaction term in the Equation 13, the interaction terms are in different order of magnitude to the main features and their coefficients are therefore not comparable. The proposal of interaction term formalisation as an alternative to 13 would be

$$\mathbf{X_{1:2}} = [\mathbf{1} \;\; \mathbf{X_1^{1/2}}] * [\mathbf{1} \;\; \mathbf{X_2^{1/2}}] \tag{15}$$

where $\mathbf{X_i^{1/2}}$ indicates the element-wise square root operation,

i.e. if $\mathbf{X_i} = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{bmatrix}$, then $\mathbf{X_i^{1/2}} = \begin{bmatrix} x_{1i}^{1/2} \\ \vdots \\ x_{ni}^{1/2} \end{bmatrix}$

5.4 Issues with binarisation of gene expression

Several methods require either the input or the response data to categorical, e.g. ANOVA requires the input vector to be discretised, while logistic regression requires the response vector to be discretised. However, both gene expression and drug sensitivity were measured as continuous values, appropriate methods were required for discretisation. ANOVA results in Chapter 4 was done by considering top 10% of expression level in all cell lines as 'sensitive' and the bottom 10% as 'resistant'. While this might ensure the difference of two groups of cell lines, 80% of the cell lines were discarded and the criterion does not truly reflect the actual distribution of the data, and this could lead to fault results. Probabilistic-based clustering methods such as Gaussian mixture model cannot be applied as it assumes that the data follows a particular distribution. The unsupervised and simple method such as k-means clustering algorithm can be the improvement for further analysis ($k = 3$). The cell lines were clustered into three groups, and the middle group are considered ambiguous and were not assigned to any cluster. This method would reduce the number of cell

lines discarded, while clustering the cell lines with taken their statistical information into account. Preliminary analysis of gene expression data with different probability densities is shown below in Figure 16.

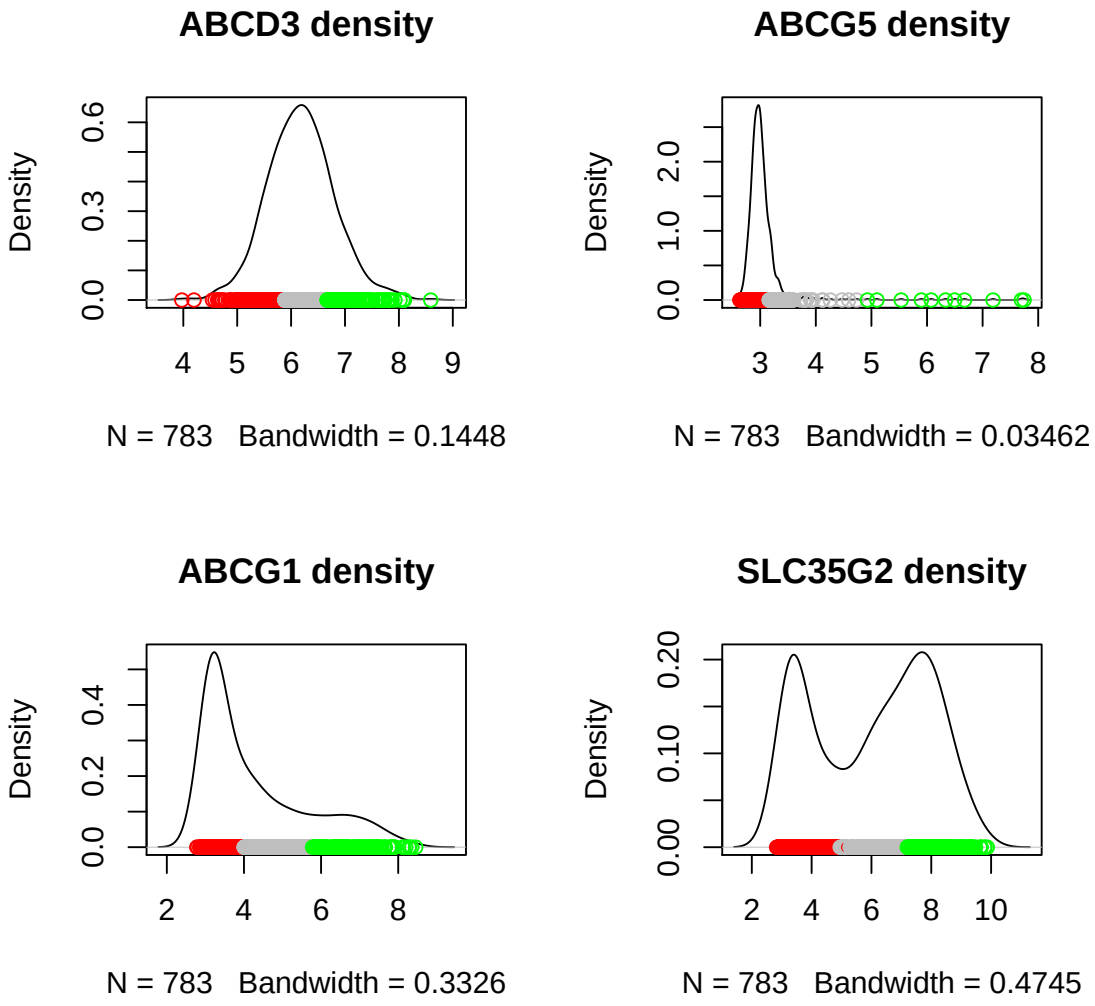


Figure 16: The density estimation of 4 genes is shown to show the behaviour of k-mean clustering algorithm when assigning $k = 3$ with Gaussian kernel applied. Data points displayed in red are labelled as 'not expressed' while data points labelled in green as 'expressed'. The data points in grey area are discarded from the analysis.

This method, however, can suffer from the imbalanced number of cell lines in each cluster. This can happen when a gene has extremely high expression level in a few cell lines, k-mean clustering tends to cluster those few cell lines in the same cluster.

6 Summary

Cellular transporters are gateways that allow small molecules, including drugs, to enter the cells. Several studies have already shown that drug efficacy can largely be determined by expression levels of particular transporters [18] [19] [6]. Evidences suggesting that toxicity and efficacy issues, which are two major concerns in drug discovery, are likely due to the lack of understanding of drug absorption and disposition [39], which can be highly related to functional roles of transporters. Nonetheless, the implication of transporter functions to cancer drug sensitivity has never been studied systematically. Solute carriers (SLC) are the second largest group of membrane transporters comprising over 400 genes, yet they are the most neglected family in the human genome, judging by the number of publications for each member of all protein families [8]. This thesis therefore aims to computationally identify putative associations between cancer cellular expression profiles of transporters and cancer drugs, as well as take a closer look to their endogenous roles in cancer tissues.

This thesis exploited the recent availability of the Genomics of Drug Sensitivity (GDSC) portal [9] [23], the largest pharmacogenomic database in cell lines. The data comprises gene expression from microarray experiments for over 15,000 genes, with the clinically significant genetic informations i.e. mutation and copy number variation, known as cancer functional events (CFEs). Drug sensitivity screening for 265 compounds in 1,001 cell lines were used to identify drug-gene associations.

The first part of the thesis focused on the landscape of transporter gene expression profiles at tissue level. Transporters that are differentially expressed in particular tissues have been identified using moderated t-test and a number of literature support for known cases were found, and novel cases were listed. Furthermore, matrix factorisation-based data integration techniques were used to integrate the shared information among different genetic data, resulting in a joint latent variable matrix that represents tissue identities by using transporter information. It was found that the identity of breast, skin, and large intestine cell lines can be recapitulated using transporter information, while most other tissues share similar transporter profiles. This information can be useful for identifications of endogenous roles of transporters in tumour progression in particular tissues. Further analyses of cell line - patient agreements of transporter profiles, which is a proxy to clinical implications, are also needed to be done.

Additional to the landscape of transporters, their effect on drug sensitivity is a central goal of this thesis. Several methods have been employed to capture these associations. Known cases such as SLC35F2 and YM155 appear as a top hit in all methods, reflecting the reliability of association methods used. SLC6A12 and DMOG is also another association pair identified in this study that appears consistently in all methods. With different sensitivity level and limitations, the other association pairs are found in not all methods, and those that can be captured by multiple methods were listed as proposed candidates for experimental validation. Furthermore, the

analyses recapitulate the known functions of ABCB1 as a multidrug resistance transporter, while providing a novel insight on the effect of SLC35F2 as multidrug sensitivity transporter.

However, the effect of one drug can be mediated via multiple transporters. The effects can either be (1) cooperativity: two transporters are needed for a drug to reach the target, or (2) redundancy: drugs can be mediated by one of the transporters independently. This question cannot be addressed by the association identification from the previous section, and therefore a first-order interaction learning method based on linear model with hierarchical group-lasso were employed to identify the transporter interactions. This is a novel analysis that can computationally suggest the joint function of transporters in drug action and the most significant pairs were listed for future experimental validation.

References

1. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32,** 40–51. ISSN: 1087-0156 (2014).
2. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov* **11,** 191–200. ISSN: 1474-1776 (2012).
3. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat Rev Drug Discov* **13,** 419–431 (2014).
4. Li, Q. & Shu, Y. Role of solute carriers in response to anticancer drugs. *Mol Cell Ther* **2,** 15. ISSN: 2052-8426 (2014).
5. Glavinas, H., Krajcsi, P., Cserepes, J. & Sarkadi, B. The role of ABC transporters in drug resistance, metabolism and toxicity. *Current drug delivery* **1,** 27–42 (2004).
6. Dobson, P. & Kell, D. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nature Reviews Drug Discovery*. doi:10.1038/nrd2438 (2008).
7. Kell, D. B. Implications of endogenous roles of transporters for drug discovery: hitchhiking and metabolite-likeness. *Nat Rev Drug Discov* **15,** 143. ISSN: 1474-1776 (2016).
8. César-Razquin, A. *et al.* A Call for Systematic Research on Solute Carriers. *Cell* **162,** 478–487. ISSN: 0092-8674 (2015).
9. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. ISSN: 0092-8674. doi:10.1016/j.cell.2016.06.017 (2016).
10. Stransky, N. *et al.* Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528,** 84–7. ISSN: 0028-0836 (2015).
11. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47 (2015).
12. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912 (2009).
13. Shen, R. *iCluster: Integrative clustering of multiple genomic data types* R package version 2.1.0 (2012). <https://CRAN.R-project.org/package=iCluster>.
14. Lim, M. & Hastie, T. Learning interactions through hierarchical group-lasso regularization (2013).
15. Nigam, S. K. What do drug transporters really do? *Nat Rev Drug Discov* **14,** 29–44. ISSN: 1474-1776 (2015).

16. Hediger, M. A. *et al.* The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. **447,** 465–468 (2004).

17. Lin, L., Yee, S. W., Kim, R. B. & Giacomini, K. M. SLC transporters as therapeutic targets: emerging opportunities. *Nat Rev Drug Discov* **14,** 543–60. ISSN: 1474-1776 (2015).

18. Winter, G. E. *et al.* The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. *Nat. Chem. Biol.* **10,** 768–73. ISSN: 1552-4450 (2014).

19. Rees, M. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* **12,** 109–116. ISSN: 1552-4450 (2016).

20. O Hagan, S., Swainston, N., Handl, J. & Kell, D. B. A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* **11,** 323–339. ISSN: 1573-3882 (2015).

21. Fletcher, J. I., Haber, M., Henderson, M. J. & Norris, M. D. ABC transporters in cancer: more than just drug efflux pumps. *Nature Reviews Cancer* **10,** 147–156 (2010).

22. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607. ISSN: 0028-0836 (2012).

23. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).

24. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *arXiv preprint arXiv:1602.08678* (2016).

25. Klami, A., Virtanen, S., Leppäaho, E. & Kaski, S. Group factor analysis. *IEEE transactions on neural networks and learning systems* **26,** 2136–2147 (2015).

26. Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. & Wild, D. L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28,** 3290–3297 (2012).

27. Bishop, C. M. Pattern recognition (2006).

28. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. **67,** 301–320 (2005).

29. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1–22 (2010).

30. Ruczinski, I., Kooperberg, C. & Michael, L. Logic Regression. *J Comput Graph Stat* **12,** 475–511. ISSN: 1061-8600 (2003).

31. Zhu, K. *et al.* p53 induces TAP1 and enhances the transport of MHC class I peptides. *Oncogene* **18** (1999).

32. Raffaghello, L. *et al.* Multiple defects of the antigen-processing machinery components in human neuroblastoma: immunotherapeutic implications. *Oncogene* **24,** 4634–4644 (2005).

33. Corrias, M. *et al.* Lack of HLA-class I antigens in human neuroblastoma cells: analysis of its relationship to TAP and tapasin expression. *Tissue antigens* **57,** 110–117 (2001).

34. Chen, H. L. *et al.* A functionally defective allele of TAP1 results in loss of MHC class I antigen presentation in a human lung cancer. *Nature genetics* **13,** 210–213 (1996).

35. Fernandez, L. *et al.* SLC45A2: a novel malignant melanoma-associated gene. *Human mutation* **29,** 1161–1167 (2008).

36. Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310,** 1782–1786 (2005).

37. Mallick, C. B. *et al.* The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet* **9,** e1003912 (2013).

38. Dixon, S. J. *et al.* Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* **149,** 1060–1072 (2012).

39. Consortium, I. *et al.* Membrane transporters in drug development. *Nat Rev Drug Discov* **9,** 215–236. ISSN: 1474-1776 (2010).

A Appendix

A.1 TCGA tissue label abbreviations

Table A1: Tissue label abbreviation used in TCGA

Tissue	**Abbreviation**
Acute Myeloid Leukemia	LAML
Adrenocortical carcinoma	ACC
Bladder Urothelial Carcinoma	BLCA
Brain Lower Grade Glioma	LGG
Breast invasive carcinoma	BRCA
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC
Cholangiocarcinoma	CHOL
Colon adenocarcinoma	COAD
Esophageal carcinoma	ESCA
FFPE Pilot Phase II	FPPP
Glioblastoma multiforme	GBM
Head and Neck squamous cell carcinoma	HNSC
Kidney Chromophobe	KICH
Kidney renal clear cell carcinoma	KIRC
Kidney renal papillary cell carcinoma	KIRP
Liver hepatocellular carcinoma	LIHC
Lung adenocarcinoma	LUAD
Lung squamous cell carcinoma	LUSC
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC
Mesothelioma	MESO
Ovarian serous cystadenocarcinoma	OV
Pancreatic adenocarcinoma	PAAD
Pheochromocytoma and Paraganglioma	PCPG
Prostate adenocarcinoma	PRAD
Rectum adenocarcinoma	READ
Sarcoma	SARC
Skin Cutaneous Melanoma	SKCM
Stomach adenocarcinoma	STAD
Testicular Germ Cell Tumors	TGCT
Thymoma	THYM
Thyroid carcinoma	THCA
Uterine Carcinosarcoma	UCS
Uterine Corpus Endometrial Carcinoma	UCEC
Uveal Melanoma	UVM

A.2 Tissue-specific transporters

This table provide the list of significant tissue specificity of SLC and ABC transporters, corresponding to Figure 8.

Solute Carriers (SLC)

gene	tissue	logFC	Avexpr	t	P.Value
SLC52A1	HNSC	1.112	3.502	14.1840	9.8E-41
SLC31A2	HNSC	1.913	4.516	13.3448	1.1E-36
SLC2A9	HNSC	1.739	3.444	11.8705	5.8E-30
SLC47A2	HNSC	1.326	3.571	10.0933	1.4E-22
RHCG	HNSC	0.718	3.232	10.0410	2.2E-22
SLC52A1	ESCA	1.025	3.502	11.5693	1.2E-28
SLC22A16	LAML	2.004	3.323	19.8926	2.4E-71
SLC39A3	LAML	1.054	3.895	12.6195	2.6E-33
SLC2A5	LAML	1.292	3.244	11.6203	7.2E-29
SLC22A15	LAML	1.248	3.322	11.4462	4.0E-28
SLC26A8	LAML	0.448	3.004	11.0963	1.2E-26
SLC6A16	DLBC	0.864	3.055	16.4065	4.3E-52
SLC2A5	DLBC	1.278	3.244	12.6351	2.2E-33
SLC35D2	DLBC	-2.597	6.420	-9.9700	4.2E-22
SLC43A2	DLBC	0.385	3.376	9.7064	4.3E-21
SLC25A27	DLBC	1.269	3.661	9.6144	9.7E-21
RHAG	LCML	3.398	3.263	12.9713	6.2E-35
SLC9A9	MM	1.182	3.212	11.8110	1.1E-29
SLC38A5	MM	3.264	3.956	10.1953	5.6E-23
SLC25A42	MM	0.768	3.453	9.7111	4.2E-21
SLC7A3	ALL	1.900	3.099	13.1035	1.5E-35
SLC25A42	ALL	0.685	3.453	9.9969	3.3E-22
SLC24A3	BRCA	2.452	3.289	15.1964	8.5E-46
SLC16A1	BRCA	-1.713	8.157	-10.1900	5.9E-23
SLC5A1	COREAD	1.586	3.183	17.9526	1.9E-60
SLC6A20	COREAD	1.588	3.376	17.7170	3.7E-59
SLC12A2	COREAD	2.735	5.455	16.0636	2.8E-50
SLC26A3	COREAD	1.750	2.995	13.6787	2.8E-38
SLC44A4	COREAD	1.937	3.668	10.8807	9.6E-26
SLC52A3	COREAD	0.559	3.619	10.7259	4.1E-25
SLC9A2	COREAD	0.933	3.257	10.6948	5.6E-25
SLC39A5	COREAD	0.908	4.047	10.2919	2.3E-23
SLC22A9	LIHC	1.110	3.001	14.2895	3.0E-41
SLC47A1	LIHC	1.019	3.312	9.8115	1.7E-21
SLC4A4	KIRC	1.219	2.988	18.5752	6.6E-64
SLC17A3	KIRC	1.141	3.272	15.2455	4.8E-46

SLC17A1	KIRC	0.860	3.107	12.9898	5.1E-35
SLC41A2	KIRC	1.622	3.948	12.8537	2.2E-34
SLC16A12	KIRC	1.650	2.965	10.1332	9.8E-23
SLC7A14	SCLC	0.982	3.150	13.6103	5.9E-38
SLC36A4	SCLC	2.083	6.056	12.9740	6.0E-35
SLC25A43	SCLC	-1.407	4.652	-12.6076	2.9E-33
SLC17A6	SCLC	0.928	3.077	11.4429	4.2E-28
SLC35D3	SCLC	1.090	3.136	11.3002	1.7E-27
SLC10A3	SCLC	-1.067	5.225	-10.1070	1.2E-22
SLC35F4	SCLC	0.721	3.082	9.8172	1.6E-21
SLC14A1	GBM	2.093	3.273	10.2755	2.7E-23
SLC6A2	NB	2.342	3.292	29.2843	3.3E-127
SLC10A4	NB	5.284	3.558	24.9132	6.8E-101
SLC4A8	NB	1.256	3.310	17.7105	4.0E-59
SLC26A10	NB	2.048	3.488	16.4052	4.4E-52
SLC18A1	NB	1.673	3.494	13.2522	3.0E-36
SLC8A3	NB	1.920	3.212	12.6501	1.9E-33
SLC6A15	NB	3.184	4.111	11.7866	1.4E-29
SLC29A4	NB	0.519	3.438	11.3125	1.5E-27
SLC8A1	NB	0.990	3.412	10.9208	6.6E-26
SLC45A2	SKCM	3.499	3.428	36.5639	2.3E-170
SLC24A5	SKCM	4.455	3.791	34.2009	1.5E-156
SLC5A4	SKCM	1.604	3.311	21.2039	6.4E-79
SLC16A6	SKCM	3.259	3.775	20.0066	5.3E-72
SLC23A2	SKCM	0.879	3.652	18.5261	1.2E-63
SLC38A1	SKCM	-3.130	10.107	-16.4763	1.8E-52
SLC1A4	SKCM	2.078	4.820	15.7776	8.6E-49
SLC35F1	SKCM	1.291	3.411	14.2512	4.6E-41
SLC6A15	SKCM	2.567	4.111	12.6728	1.5E-33
SLC35B4	SKCM	1.591	5.646	12.4775	1.1E-32
SLC35B2	SKCM	1.020	6.439	12.2971	7.4E-32
SLC6A8	SKCM	1.791	6.578	11.5757	1.1E-28
SLC27A1	SKCM	0.303	3.236	11.5395	1.6E-28
SLC37A3	SKCM	0.737	4.182	10.7844	2.4E-25
SLC19A2	SKCM	1.435	5.141	10.6884	5.9E-25
SLC26A4	SKCM	0.776	2.996	10.6679	7.2E-25
SLC27A3	SKCM	0.909	3.642	9.9013	7.8E-22
SLC26A2	SKCM	1.894	6.528	9.7312	3.5E-21
SLC34A2	OV	2.094	3.154	13.9837	9.4E-40

ATP-binding cassettes (ABC)

gene	tissue	logFC	AveExpr	t	P.Value
ABCB10	LAML	1.429	7.494	10.9721	4.0E-26
ABCD2	MM	0.611	2.798	9.9638	4.5E-22
ABCC11	BRCA	0.689	3.151	14.5259	2.0E-42
CFTR	COREAD	2.537	3.439	16.9268	7.2E-55
ABCA5	SCLC	1.209	3.542	13.2399	3.4E-36
TAP1	SCLC	-1.925	7.456	-10.0271	2.5E-22
TAP1	NB	-3.282	7.456	-12.4167	2.2E-32
ABCB5	SKCM	0.555	3.007	18.5492	9.3E-64
ABCB4	SKCM	1.226	3.305	12.4116	2.3E-32
ABCG5	SKCM	0.625	3.064	10.0183	2.7E-22